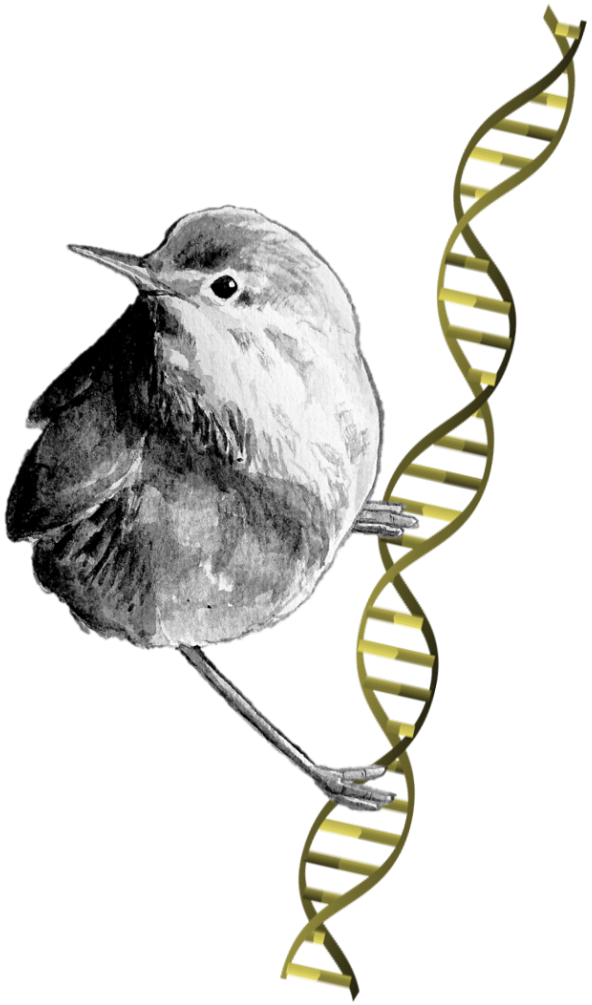
Escaping the genetic costs of a range expansion?

The case of a migratory passerine (*Acrocephalus scirpaceus*)



Nora Bergman

Master's thesis

Master's Programme in Ecology and Evolutionary Biology

Faculty of Biological and Environmental Sciences

University of Helsinki

May 2021

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Abstract

Faculty:	Faculty of Biological and Environmental Sciences
Degree programme:	Master's Programme in Ecology and Evolutionary Biology
Author:	Nora Bergman
Title:	Escaping the genetic costs of a range expansion? The case of a migratory passerine (*Acrocephalus scirpaceus*)
Level:	Master's thesis
Month and year:	5/2021
Number of pages:	43 + 8 (appendices)
Keywords:	Genetic diversity, Passeriformes, population genomics, population structure, RAD-seq, range expansion
Supervisor or supervisors:	Dr. Katja Rönkä, Dr. Rose Thorogood, Dr. Perttu Seppä
Where deposited:	HELDA – Digital Repository of the University of Helsinki

Abstract:

Rapid environmental changes over the last 100 years have led to substantial range shifts across taxonomic groups. Understanding what facilitates successful shifts is important for predicting ecological consequences and planning efficient conservation actions. Interestingly, the very process of range expansion can affect the success of the shift by causing genetic changes in the expanding populations. Theory predicts that without sufficient gene flow, repeated founder events and strong genetic drift can result in allele frequency gradients and loss of genetic diversity along the expansion axis. Empirical studies testing these expectations in environment-driven range shifts are still relatively scarce, and how range expansions affect genetics in highly mobile species remains unclear.

In this study, I investigated the genetic consequences of a recent range expansion in a long-distance migratory passerine, the reed warbler (*Acrocephalus scirpaceus*). Utilizing genome-wide data from restriction site-associated DNA sequencing (RAD-seq), I studied whether the expansion was reflected in either population structure or genetic diversity of the recently established Finnish range edge population. Despite philopatry and genetic differentiation to the range core populations, principal component analysis (PCA) and a model-based Bayesian clustering approach (fineRADstructure) revealed a lack of spatial population structure along a putative colonization route. Levels of genetic diversity, based on expected heterozygosity, nucleotide diversity, and private allele count, were found to be very similar between range edge (Finland) and range core (Central Europe). The results likely indicate high levels of gene flow both within the new population and across greater spatial distances during or after the range expansion. Due to a detected sequencing batch effect, however, the exact diversity estimates must be considered preliminary.

These findings suggest that species with high enough dispersal propensity may escape the predicted genetic costs of range expansions, retaining high levels of genetic variation at range margins. This study provides valuable insights for understanding range shifts in mobile taxa, and highlights the need to investigate further the traits of species that enable the preservation of evolutionary potential during range shifts.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiivistelmä

Tiedekunta:	Bio- ja ympäristötieteellinen tiedekunta
Koulutusohjelma:	Ekologian ja evoluutiobiologian maisteriohjelma
Tekijä:	Nora Bergman
Työn nimi:	Voiko levittäytymisen geneettiset kustannukset välttää? Esimerkki rytikerttusen (*Acrocephalus scirpaceus*) levinneisyysalueen reunalta
Työn laji:	Maisterintutkielma
Kuukausi ja vuosi:	5/2021
Sivumäärä:	43 + 8 (liitteet)
Avainsanat:	Geneettinen monimuotoisuus, levittäytyminen, populaatiogenomiikka, populaatiorakenne, RAD-sekvensointi, varpuslinnut
Ohjaaja tai ohjaajat:	FT Katja Rönkä, FT Rose Thorogood, FT Perttu Seppä
Säilytyspaikka:	HELDA – Helsingin yliopiston digitaalinen arkisto
Tiivistelmä:	

Lukuisten eliölajien levinneisyysalueet ovat siirtyneet kuluneen vuosisadan aikana ympäristönmuutosten seurauksena. Jotta voidaan ennakoida siirtymisen ekologisia vaikutuksia ja suunnitella tehokkaita suojelutoimia, on tärkeää ymmärtää levittäytymisen onnistumiseen vaikuttavia tekijöitä. Uusien alueiden asuttaminen voi vaikuttaa suoraan levittäytymismenestykseen aiheuttamalla geneettisiä muutoksia levittäytyvässä populaatiossa. Toistuvien perustajavaikutusten ja geneettisen ajautumisen odotetaan aiheuttavan esimerkiksi alleelifrekvenssien muuttumista ja geneettisen monimuotoisuuden hupenemista levittäytymisen etenemissuunnassa, etenkin geenivirran ollessa vähäistä. Suhteellisen harvat empiiriset tutkimukset ovat kuitenkaan testanneet näitä oletuksia, ja levittäytymisen vaikutukset varsinkin hyvin liikkuvaisiin lajeihin tunnetaan huonosti.

Tutkielmassani selvitin äskettäisen levittäytymisen geneettisiä vaikutuksia varpuslintuihin kuuluvan rytikerttusen (*Acrocephalus scirpaceus*) levinneisyyden pohjoisrajalla. Genominlaajuisen RAD (*restriction site-associated DNA*) -sekvensointiaineiston avulla tutkin, ilmeneekö levittäytyminen äskettäin perustetun populaation geneettisessä rakenteessa tai monimuotoisuudessa. Huolimatta lajin pesimäpaikkauskollisuudesta, reunapopulaatiossa ei havaittu oletettua levittäytymisreittiä vastaavaa populaatiorakennetta. Geneettisen monimuotoisuuden määrä, perustuen nukleotidien monimuotoisuuteen (π), odotettuun heterotsygotia-asteeseen (H_e) ja privaattialleelien lukumäärään, osoittautui hyvin samansuuruiseksi levinneisyyden reunalla (Suomi) ja ydinalueella (Keski-Eurooppa). Tulokset ovat mitä luultavimmin osoitus runsaasta geenivirrasta sekä reunapopulaatiossa että maantieteellisesti laajemmilla alueilla. Vaikka reuna- ja ydinpopulaatioiden välillä on geneettistä eriytymistä, valtaosa muuntelusta vaikuttaa säilyneen levittäytymisen aikana. Tarkat arviot geneettisen monimuotoisuuden määrästä ovat kuitenkin alustavia sekvensointierien välillä havaitun teknisen eron vuoksi.

Nämä löydökset viittaavat siihen, että riittävän liikkuvaiset lajit voivat välttää levittäytymisen ennustetut geneettiset kustannukset ja ylläpitää runsasta geneettistä muuntelua myös levinneisyysalueensa reunoilla. Tämä tutkielma tarjoaa tärkeän esimerkin siitä, millaisia vaikutuksia levinneisyysalueiden siirtymisellä on odotettavissa liikkuvaisimpiin lajeihin. Se myös korostaa tarvetta tutkia tarkemmin ominaisuuksia, jotka mahdollistavat evolutiivisen potentiaalin säilymisen muuttuvissa elinympäristöissä.

Table of contents

1. INTRODUCTION

1.1 Range shifts

Species' distributions are anything but static. Range boundaries change and shift in the course of time, usually either expanding or contracting at their margins. Changes take place both over long time periods (e.g. in response to glacial episodes) and over much shorter times (reviewed in Brown et al. 1996). The range of a species can be thought of as its ecological niche in space (Sexton et al. 2009), and therefore changes in the range tend to reflect either niche evolution (meaning the evolutionary adaptation to novel conditions) or spatial tracking of the existing niche (Gaston 1998, Pfenninger et al. 2007). These processes allow species to follow suitable conditions that shift in space, or to expand their range into entirely new areas and ecosystems. While range shifts that occur during long temporal scales have been recognized and studied for a long time (e.g. using the fossil record), there is currently increasing interest and urgency to study also the more rapidly occurring shifts. Anthropogenic activities are altering ecosystems at accelerating rates (e.g. Pereira et al. 2010, Waters et al. 2016, Pecl et al. 2017), and range shifts are one possibility for species to respond to many of the changes. Global meta-analyses have already documented rapid range shifts across different taxa, at a median rate of 16.9 km per decade towards higher latitudes and 11.0 meters per decade towards higher elevations, largely matching the documented global warming (Chen et al. 2011, IPCC 2014).

The need to understand range shifts under the current environmental change is important for two major reasons. First, range shifts may enable species to persist in changing conditions, and therefore predicting and possibly facilitating these shifts is essential for efficient conservation (e.g. Hannah et al. 2002, Alagador et al. 2016). Second, range shifts can have wider ecological consequences, altering community structure and ecosystem processes. Both species that track their environmental niche by range expansion, and introduced species that are transported to new ecosystems by humans, have been documented to impact the recipient communities (reviewed in Wallingford et al. 2020). Introduced species are especially likely to become invasive and have negative impacts, potentially because they often lack a shared evolutionary history with the recipient community (e.g. Jeschke & Strayer 2005, Simberloff et al. 2012). However, also other range shifts will almost inevitably lead to changes or disruptions in species interactions, as communities are very unlikely to shift as whole (Tylianakis et al. 2010; Wallingford et al. 2020).

In all these considerations, the central questions are why certain species respond to environmental changes by shifting their range while others do not, and why some succeed in colonizing new areas while others fail. Meta-analyses suggest that successful range shifts are combinations of many ecological and evolutionary

factors (e.g. Angert et al. 2011, MacLean & Beissinger 2017). Along with environmental aspects (e.g. environmental gradients and their steepness; Pigot et al. 2010, Polechová 2018) and species' traits (e.g. generalism; MacLean & Beissinger 2017), the ability to adapt to novel conditions has been largely associated with colonization and establishment success in new areas (e.g. Crawford & Whitney 2010, Szűcs et al. 2014; Wennersten & Forsman 2012, Wallingford et al. 2020). Interestingly, maintaining high adaptive potential while going through a range expansion is not self-evident at all. A range expansion itself can greatly affect the genetic composition and therefore also the adaptive potential of the colonizing population (e.g. Pujol & Pannell 2008, Colautti et al. 2010; Excoffier et al. 2009), which is also important to account for when predicting the dynamics of range shifts (e.g. Fordham et al. 2014).

1.2 Range expansions can lead to significant genetic changes in the expanding population

There is a variety of population genetic processes that can change the course of evolution in spatially expanding populations. These processes can for instance reduce genetic diversity in the newly colonized areas, generate allele frequency gradients, or promote the spread of rare or even deleterious variants at the range edge (reviewed in Excoffier et al. 2009). There can also be introgression with local species (e.g. Garcia-Elfring et al. 2017; Excoffier et al. 2009) and selection for locally adaptive or expansion-facilitating traits (e.g. Liebl & Martin 2012, Savolainen et al. 2013, Lombaert et al. 2014). The following sections will focus on the processes that affect population structure and genetic diversity in spatially expanding populations, comparing the theoretical expectations with existing empirical observations.

1.2.1 Genetic consequences of a range expansion

One of the most important evolutionary forces during range expansions is genetic drift (i.e. change in allele frequencies due to chance). Colonization of new areas is often characterized by repeated founder effects at the leading range edge, caused by a limited number of founder individuals, and small effective population sizes (Welles & Dlugosch 2019). Strong genetic drift due to these factors may lead to reduced genetic diversity accompanied with increased genetic structure and differentiation at the range front, especially in the absence of sufficient gene flow from other parts of the range (Austerlitz et al. 1997; reviewed in Excoffier et al. 2009). Loss of genetic variety can reduce the adaptive potential in the newly colonized areas, which may slow down or even halt the expansion (e.g. Pujol & Pannell 2008, Szűcs et al. 2017, Nadeau &

Urban 2019). Another typical consequence of the sequential founder effects is the so-called allelic surfing, as certain, even rare or deleterious alleles from the founder population rise to high frequencies or become fixed in the range edge population due to a spatial equivalent of genetic drift. The alleles can be imagined to "surf" on the expansion wave along with the lineages that have a high breeding success in the newly colonized areas (Klopfstein et al. 2006).

Empirical studies on recent or currently ongoing expansions are still relatively scarce, but genetic signatures of range expansion, matching the theoretical predictions, have been shown to appear also in wild populations. Reduced genetic diversity along the expansion axis has been reported for instance in populations of balsam poplars (Keller et al. 2010), flying squirrels (Garroway et al. 2011), bank voles (White et al. 2013), white-footed mice (Garcia-Elfring et al. 2017), and coral symbionts (Grupstra et al. 2017). Genetic structuring or differentiation, reflecting the colonization routes, has been detected in e.g. damselflies (Swaegers et al. 2013), monarch butterflies (Pierce et al. 2014), hazel grouses (Rózsa et al. 2016) and coyotes (Heppenheimer et al. 2018). An increase in putatively deleterious mutations during the expansion has been reported in e.g. a species of Asteraceae, *Leontodon longirostris* (de Pedro et al. 2021), and patterns of allele surfing at nearly all studied loci were found in the common wall lizard (Gassert et al. 2013).

However, many studies have not detected these expected patterns. For instance, there were no signs of reduced genetic diversity in coyotes despite clear population structure, likely due to hybridization with other *Canis* species (Heppenheimer et al. 2018). In a sister species complex of two passerines, the melodious warbler and the icterine warbler, no genetic structure or changes in diversity were detected along either expanding or receding range edges (Engler et al. 2015). In absence of signs of hybridization, the lack of genetic signatures was hypothesized to result from high mobility and long-distance dispersal (LDD) events. Similarly, surprisingly high genetic diversity was observed after an invasion of European starlings in Africa, suggested to result from frequent LDD events. However, the possibility of multiple introductions could not be ruled out (Berthouly-Salazar et al. 2013). In an invasive lionfish with high dispersal capabilities, population structure could be detected along the invasion pathway, but no reduction in genetic diversity (Bors et al. 2019). Some studies have even been able to document the breakdown of the genetic patterns in action, observing an increase in genetic diversity and decrease in the degree of genetic structure after the initial colonization. In brown bears, these changes were detected in just 1.5 generations, resulting from e.g. substantial immigration from neighboring populations (Hagen et al. 2015). There is great variation in the genetic patterns after range expansions, but there is also great variation in spatial and temporal scales of the studied range expansions, as well as the traits of the expanding species and the environment. How does this interplay of different factors shape the genetic consequences of range expansions? In what time frames can different factors be expected to attenuate the initial genetic patterns?

1.2.2 Important factors in shaping case-specific genetic patterns

Theoretical studies have aimed to identify the most important factors that affect the intensity and duration of the genetic consequences of range expansions. Especially the speed of the expansion and factors that mitigate loss of genetic diversity have been found to be central (reviewed in Excoffier et al. 2009). The speed of the expansion can be slowed down by environmental heterogeneity and the requirement for local adaptation (e.g. Wegmann et al. 2006, Gilbert et al. 2017), the presence of an Allee effect (i.e. a lower per capita growth rate at low population densities) (Roques et al. 2012), or interspecific competition (e.g. Skellam 1951, Legault et al. 2020). A slow expansion may retain genetic diversity and reduce the genetic expansion load at a longer timescale, as it gives gene flow from other parts of the range a chance to "catch up". The expansion speed can in turn be accelerated by the mixture of individuals from multiple source populations (Wagner et al. 2017), or by frequent long-distance dispersal events. However, fast range expansions do not necessarily lead to reduced diversity: LDD is an example of a phenomenon that both speeds up the expansion and preserves genetic diversity in the new populations (e.g. Kawecki 2000, Berthouly-Salazar et al. 2013). Factors that mitigate the loss of genetic diversity include for instance higher growth rate, which allows the population to better recover after bottlenecks, higher migration rates, which can preserve and restore genetic diversity at range front (Austerlitz et al. 1997), and potentially also interspecific introgression, which may provide beneficial genetic variation for adaptation in the new environment (Pfennig et al. 2016). A reduced number of founder individuals has been found to accelerate the loss of genetic diversity (e.g. Whitlock & McCauley 1990). Diverse processes affect the dynamics of the advancing expansion, and the demographic history of the expanding populations, in terms of e.g. standing genetic variation before the expansion, should not be overlooked either.

1.3 A local newcomer: Eurasian reed warbler (*Acrocephalus scirpaceus*)

1.3.1 Expanding the species' northern range edge to Finland

The Eurasian reed warbler *Acrocephalus scirpaceus* (hereafter "reed warbler") is a geographically widespread long-distance migratory passerine, mainly breeding in stands of common reed *Phragmites australis* (hereafter "reed") in Eurasia and Northern Africa, and migrating to sub-Saharan Africa for wintering (Cramp & Brooks 1992). Reed warbler is an example of a species that has recently undergone a

range expansion at its northern range limit. The range of the species has been expanding in Europe towards north over the documented history, advancing through Denmark to Sweden in the latter half of the 19th century (Løppenthin 1967: cited in Avilés et al. 2006, Järvinen & Ulfstrand 1980), and reaching Finland and Norway in the first half of the 20th century (Røed 1994, Valkama et al. 2011: Finnish Breeding Bird Atlas) (Fig. 1). The first reported observation of the species in Finland was made on the Åland islands in 1926 (Leivo 1937), while it has been confirmed that the first reed warblers arrived already some years earlier (Wikström 1945). The size of the Finnish reed warbler population has since expanded rather quickly, from around 500 nesting pairs in the 1950's to the estimated 20 000–30 000 pairs today (Valkama et al. 2011: Finnish Breeding Bird Atlas). The expansion was especially rapid until the 1980's, after which its advancement and the species' population size began to level. Currently, the species' main range in Finland is located at the southwestern and southern coast, most of the confirmed nestings being from this region. Reed warblers can also be found at inland waters in central and eastern Finland, as well as in coastal reed beds further up north (Valkama et al. 2011: Finnish Breeding Bird Atlas). What exactly facilitated the northward range expansion of reed warblers has not been confirmed, but a notable increase in the amount of reed bed habitat in Finland during the last century is likely to be an important factor (Koskimies 1981, Altartouri et al. 2014). The expansion of reed bed habitat probably results from land use changes and eutrophication, possibly accompanied by climatic changes (Ikonen & Hagelberg 2007, Altartouri et al. 2014).

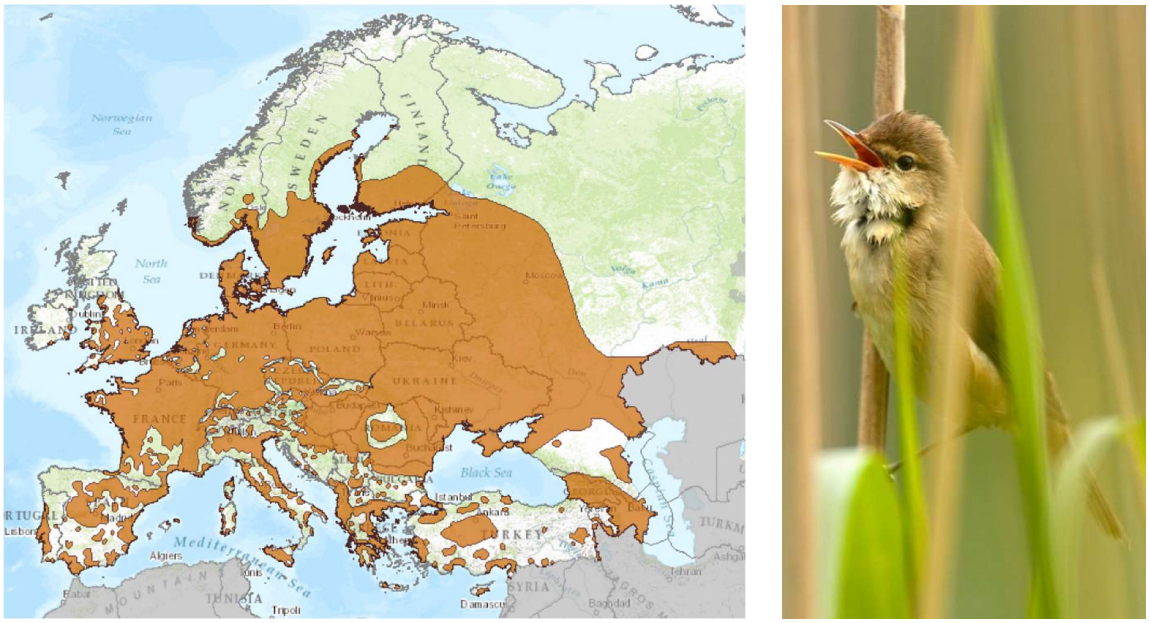


Figure 1. Left: The European breeding range of the reed warbler (shown in orange), adapted from BirdLife International 2015, European Red List of Birds. Right: A singing reed warbler in its breeding habitat. Image: Chris Romeiks / vogelart.info.

1.3.2 Dispersal-related and genetic characteristics of the reed warbler

Comparing with many other groups of organisms, the capability of flying already gives birds an advantage in their dispersal potential. Migratory species have especially high dispersal potential due of their lifestyle, which involves annual journeys between breeding and non-breeding grounds, and they also tend to disperse further than resident species (Paradis et al. 1998). High dispersal potential facilitates gene flow, consequentially affecting genetic diversity and the homogeneity among populations (Slatkin 1985). However, the capacity to disperse does not necessarily correspond to the realized dispersal patterns. Many bird species, including the reed warbler, exhibit strong fidelity to their natal location, also known as philopatry (Greenwood 1980, Paradis et al. 1998). The natal dispersal (dispersal of juveniles from birth location to their first own nesting sites) and breeding dispersal (movement between subsequent breeding sites) of reed warblers has been estimated from British ringing data, spanning the years 1909-1994 (Paradis et al. 1998). While the vast majority of individuals returned to the exact same location to breed, occasional longer dispersal events ranging from a few kilometres even up to 300 km were also documented. For breeding dispersal, the geometric mean was 2.9 km and arithmetic mean 32.4 km. The values for natal dispersal were 5.2 km and 47.0 km, respectively. This presents an intriguing background for studying range expansion in the species, examining the effect of regular but infrequent long-distance dispersal events in a relatively short timeframe.

There are three recognized subspecies of the reed warbler, of which the nominal subspecies *A. scirpaceus scirpaceus* is the focal subspecies in this study, breeding in Europe and Northern Africa (Cramp & Brooks 1992, Arbabi et al. 2014). The genetics of the reed warbler has been previously studied to examine Europe-wide population structure and gene flow patterns (Procházka et al. 2011) and resolve the phylogeography and evolutionary history of the species (Arbabi et al. 2014). Using DNA microsatellite markers, Procházka and colleagues (2011) found low levels of genetic differentiation among European populations, the greatest differentiation being between 1) Iberian and other European populations (suggesting the role of the Iberian Peninsula as one of the Pleistocene glacial refugia), 2) the different subspecies, and 3) populations on different sides of a migratory divide (i.e. geographical boundary of divergent migration directions) in Central Europe. With the resolution of DNA microsatellite markers, no population structure could be detected among the rest of the populations. With the use of mitochondrial sequences, Arbabi and colleagues (2014) found that levels of mitochondrial diversity were similar across the species' range, although the diversity in Northern European populations was not measured due to small sample sizes. These findings indicate high levels of gene flow, but also demonstrate the need for more high-resolution methods to explore the population structure in more detail.

1.4 Aims of the study

The objective of this study is to explore how a recent range expansion has genetically affected the newly founded reed warbler population at the northern range edge. I aim to answer the following questions:

1) Is the recent range expansion reflected in the genetic structure of the Finnish reed warbler population? I expect to detect genetic structuring from west to east along the southern coast of Finland, caused by successive founder effects or allelic surfing at the advancing range front. The west–east gradient reflects the most likely colonization route of the species, based on the observation data from the early years of the colonization and current migratory routes suggested by recaptured ringed individuals (Fransson & Stolt 2005). An alternative hypothesis is that no genetic structure can be detected, which could be caused by e.g. stronger gene flow from areas further away from the range front, or the occurrence of sufficiently frequent LDD events despite philopatry.

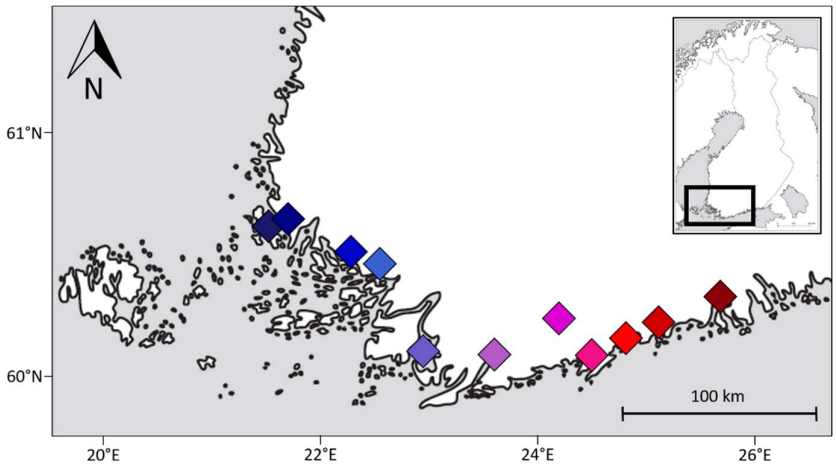
2) Is there reduced genetic diversity in the Finnish range edge population, compared to range core populations in Central Europe? I expect to find lower diversity at the range edge than at the range core due to founder effects and genetic drift, considering the short time since the colonization and the notable advancement of the range edge (> 1000 km) during the past two centuries. Alternatively, the genetic diversity might not be significantly lower at the range edge, indicating maintained diversity during the initial colonization or restored diversity after it.

To answer the study questions, I use genomic data from individuals at the range edge (Finland) and the range core (Central Europe). The samples are sequenced using Restriction-site Associated DNA sequencing (RAD-seq; Miller et al. 2007b, Baird et al. 2008), which provides high resolution for detecting genetic signals even in populations with low levels of genetic differentiation. Studying the population structure in the newly colonized area and estimating the amount of retained genetic diversity will give a better understanding about this specific range expansion, but also provide an example of the genetic effects of a range expansion in a highly mobile but philopatric species. In addition, a population genetic comprehension can be valuable when studying other aspects of reed warbler ecology and evolution, such as evolutionary or behavioural adaptation to local conditions in the recently founded population.

2. MATERIAL AND METHODS

2.1 Description of the data and study design

The data consists of 84 RAD-sequenced reed warbler individuals from Finland, 19 individuals from the Czech Republic and Slovakia, and 5 re-sequenced controls. The 84 Finnish samples are collected from reed bed sites in 11 municipalities, forming a spatial gradient along the southern coast of Finland (Fig. 2). These samples represent the range edge population of the species, and the sampling design allows the detection of any potential population structure that could result from the putative historical advancement of the range front from west to east. The 19 Czech and Slovakian samples are collected from three different sites, representing the reed warbler range core in Europe. These samples are used to compare the genetic diversity between the range core and the range edge areas, together with a geographically comparable subset of the Finnish samples (Fig. 3).

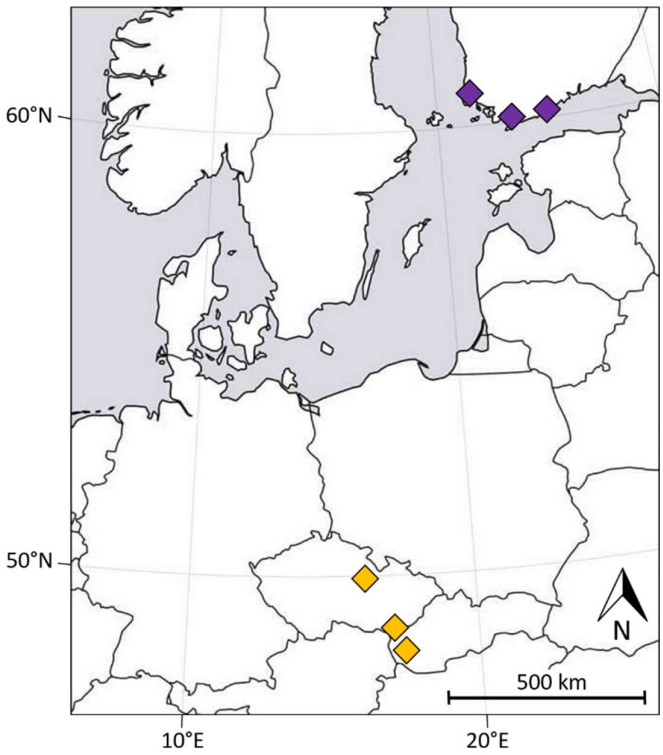


	Location	N
	Kustavi	2
	Taivassalo	4
	Raisio	1
	Kaarina	6
	Salo	4
	Raasepori	2
	Siuntio	9
	Kirkkonummi	9
	Espoo	18
	Helsinki	10
	Porvoo	19

Figure 2. Sampling locations at the southern coast of Finland. The number of sampled individuals from each municipality are shown in the table.

The five re-sequenced control samples are used to ensure the comparability between two separate sequencing batches. Due to a newly established collaboration in reed warbler research, the Central European samples were sequenced first in a single batch (Batch 1), and the Finnish samples thereafter in another batch (Batch 2). The control samples contain five individuals sequenced in both batches from the same DNA extracts. Although different high-throughput sequencing batches, including RAD-seq, are

routinely combined for analyses, a potential complication is that there can be systematic differences in the outcomes of different sequencing batches, also known as batch effects (Leek et al. 2010). The inclusion of control samples allows the detection of potential batch effects, discussed in more detail later.



Location	N	Batch
Range edge (Finland)	19	B2
Kustavi + Taivassalo	5	
Salo + Raasepori	5	
Helsinki	9	
Range core (Czech R. + Slovakia)	19	B1
Osík	5	
Luzice	5	
Trnava	9	
Controls	5	B1 + B2

Figure 3. Sampling locations for genetic diversity analysis between range edge (Finland; purple) and range core (Czech Republic and Slovakia; yellow) populations. The total number of sampled individuals, the number of individuals from each sampling location, and the batch in which the samples were sequenced are shown in the table. Batch 1 contains samples also from other European countries, and the five control samples include one re-sequenced sample from each of the following countries: the Czech Republic, France, Italy, Norway and Turkey.

2.2 Blood sample collection and DNA extraction

Blood samples from the Finnish reed warbler population were taken from 84 individual birds caught at their territories using playback song and netting, with the permissions of the Regional State Administrative Agency for Southern Finland (ESAVI/3920/2018), Southwest Finland Centre for Economic Development, Transport and the Environment (VARELY/758/2018, VARELY/799/2019) and Metsähallitus (MH1100/2018/06.06.02). The caught birds were ringed (ringing permissions via the Finnish Ringing Centre), making it possible to identify previously captured individuals. The blood samples were stored in

Queen's lysis buffer at –20 °C. The sampling was performed by Dr. Rose Thorogood's research group (University of Helsinki) in summer 2018. To learn the field methodology and to get to know the study system, I took part in similar field work in summer 2020. The samples from Central European sites were provided by the research group's collaborators at the University of Oslo. The blood samples were collected from birds close to their territories (not on migration), and stored in either buffer or ethanol at +4 °C.

Genomic DNA was extracted using the QIAGEN© DNeasy Blood & Tissue Kit (Qiagen Inc.). Before applying the manufacturer's protocol, the samples were first handled according to their storage medium (Queen's lysis buffer or ethanol). Buffer samples (125 µl) were mixed thoroughly with 75 µl Phosphate Buffered Saline (PBS) and 20 µl Proteinase K, and then incubated overnight. The ethanol samples were first treated to evaporate all ethanol from the solid blood cell pellet using a heat block. After that, the sample was thoroughly mixed with 180 µl Buffer ATL (Qiagen) and 20 µl Proteinase K, letting to incubate overnight. The following day, 4 µl RNase A was added to all samples, mixed and incubated for two minutes at room temperature. After this, the extraction kit's standard protocol for blood samples was followed starting from step 2, with only minor changes to incubation times and the number of eluates prepared per sample.

After extraction, the purity and DNA concentration of the eluates was assessed using the NanoDrop 2000 spectrophotometer (Thermo Scientific) and Qubit 4 fluorometer (Invitrogen). Only eluates with satisfactory absorbance ratios (NanoDrop: 260/280 ratio ~1.8 – 2.0; 260/230 ratio > 1.5) indicating sample purity, and high enough DNA concentration estimate (NanoDrop: > 20 ng/µl), were accepted. The DNA concentration was measured more precisely using the Qubit fluorometer, as its measurements are based on the detection of fluorescent dyes that will only bound to the target molecules (DNA), and are therefore not affected by e.g. proteins or salts. All samples were normalized to a DNA concentration of 20 ng/µl as required by the sequencing company. Finally, gel electrophoresis was used to ensure that the DNA fragments were of desired size (molecular weight > 10 kb). The DNA extraction and sample preparation were carried out by Dr. Katja Rönkä in 2018. To learn the methodology, I carried out similar DNA extractions in 2020 for 93 samples that will be used in upcoming studies of the research group.

2.3 Sample sequencing and building the RAD loci

The samples were sequenced by Floragenex, Inc. (Oregon U.S.A.) in 2019 using the original RAD-seq (restriction site-associated DNA sequencing) method. As its name implies, RAD-seq uses restriction enzymes to cut and sample the DNA at enzyme-specific restriction sites, which are spread across the genome. The genomic DNA was digested with a single restriction enzyme, followed by ligation of Floragenex RAD-seq adapters with individual library indexes, shearing of the DNA with sonication, end

repair and gel purification of the fragments, ligation of Floragenex secondary RAD-seq adapters to the library, and finally polymerase chain reactions (PCR). As an end product, corresponding sequence blocks (length 91 bp) were acquired from the cutting sites, creating a relatively randomised subset of the full genome of each sequenced individual.

The Stacks pipeline (v2.54; Catchen et al. 2013, Rochette et al. 2019) was used for building the loci from the sequence data. Stacks was developed to work with restriction enzyme-based data, such as RAD-seq, and can be used for either *de novo* (i.e. without reference genome) or reference genome based analyses. This study follows the reference-based protocol, using a recently completed genome of the reed warbler (Camilla Lo Cascio Sætre, in preparation). The protocol consists of three major stages: 1) cleaning and demultiplexing the reads, 2) assembling the reference-aligned loci according to alignment positions and calling SNPs in each sample, and 3) filtering the data, calculating population genetic statistics and exporting data in different formats. The first stage of the pipeline and aligning loci to the reference genome were performed by Dr. Katja Rönkä. I continued with the pipeline from the second stage forward, and did the data assessment, filtering and analyses.

The cleaning and demultiplexing was performed with the *process_radtags* program in Stacks. One sequencing lane simultaneously sequences multiple individuals, and demultiplexing identifies and sorts the reads using individual-specific tags, also called barcodes. The *process_radtags* program first checks that there are no errors in the barcode or the RAD cut site and corrects a number of possible mismatches, then checks the average quality score using a sliding window. The program was run with default settings: allowing for maximum one mismatch in the barcode or cut site, and using a sliding window that is 15 % of the length of the read, discarding the read if the average raw Phred score drops below 10. The numbers of removed and retained reads after completing *process_radtags* are listed in Table 1.

Table 1. The number of sequences in each sequencing batch before and after running the *process_radtags* unit. In both batches, less than 10 % of the total number of reads were discarded due to issues in barcodes, cut sites or quality.

Sequencing batch	Total sequences	Discarded due to an erroneous barcode	Discarded due to an erroneous RAD cut site	Discarded due to low quality	Retained reads
Batch 1 (Europe)	1 105 200 841	64 255 834 (5.8 %)	6 246 707 (0.6 %)	486 370 (0.0 %)	1 034 211 930 (93.6 %)
Batch 2 (Finland + controls)	705 540 255	60 483 503 (8.6 %)	3 326 578 (0.5 %)	3 237 891 (0.5 %)	638 492 283 (90.5 %)

The cleaned sequences were aligned against the reference genome using the BWA-MEM alignment algorithm (v0.7.17-r1188; Li 2013), using default parameter values. After this, the resulting BAM files were sorted and the *ref_map.pl* program in Stacks was used to assemble the loci according to the alignment positions and to call SNPs in each sample. The final step of the pipeline, the *populations* program for filtering the data and calculating population genetic statistics, was run multiple times with different sample assemblies and parameters to produce the desired output for each study question and analysis. The parameter choices for *populations* are discussed in detail in sections 2.4.1 and 2.5.

2.4 Data quality, comparability and filtering

2.4.1 Quality assessment and filtering protocol

The final filtering was carried out using the *populations* program in Stacks, but as the Stacks pipeline does not include an option for read depth filtering, the program *VCFtools* (v0.1.16; Danecek et al. 2011) was first used for creating a blacklist for *populations* for this purpose. For outputting the SNPs in Variant Call Format (VCF), the Stacks *populations* program was first run without applying filters. *VCFtools* was then used to extract a number of output statistics from the VCF file (mean depth per site and per individual, and missingness per site and per individual) in order to assess the data and to determine the depth filtering thresholds and the need to remove any individuals due to e.g. high amounts of missing data. These output files were examined and plotted using R (v4.0.0; R Core Team) using the package *tidyverse* (Wickham et al. 2019).

After this, *VCFtools* was used to filter the VCF file for minimum and maximum mean depth per site. The threshold for minimum mean depth was set to 15 to prevent false positive calls. A maximum mean depth threshold of approximately two times the mean depth across all sites was used. The maximum depth filter cuts off sites with extremely high read depths as these are likely to be mapping errors, combinations of multiple repetitive regions in the genome (O'Leary et al. 2018). At this filtering step, also the sex chromosome was removed (see section 3.1.1). The "removed-sites" output option in *VCFtools* was used to produce a list of all filtered sites. Using a custom Unix code, this list was converted into a blacklist file for *populations*. The blacklist excludes all the loci that contain one or more filtered sites. Removing the entire loci instead of single SNPs was chosen in order to maintain the integrity of haplotypes.

Finally, the *populations* program was run with analysis-specific settings. These can be found in detail under section 2.5, but an overview of the different utilized filtering options and their purposes is compiled in Table 2.

Table 2. A summary of the different data filtering options used in this study.

Program	Filter	Use	Purpose
Stacks: *populations*	-p	Minimum no. of populations a locus must be present in to be processed	Reducing missing data and providing biological control for the specific question
Stacks: *populations*	-r	Minimum % of individuals per population a locus must be present in to be processed	Reducing missing data and providing biological control for the specific question
Stacks: *populations*	--min-mac	Minimum minor allele count for a SNP to be processed (across all populations)	Excluding rare, potentially erroneous alleles
Stacks: *populations*	--max-obs-het	Maximum observed heterozygosity for a SNP to be processed (across all populations)	Excluding potentially erroneously merged sites with very high levels of heterozygosity
Stacks: *populations*	--write-random-snp	Restricts data analysis to one random SNP per locus	Removing tightly linked SNPs from analyses that assume unlinked markers
Stacks: *populations*	--blacklist	A list of loci to not be processed	(In this study: removing all loci with any number of sites excluded by the VCFtools filters below)
VCFtools	--min-meanDP	Minimum mean depth value for a site to be included	Removing low-depth sites to prevent false positive SNP calls
VCFtools	--max-meanDP	Maximum mean depth value for a site to be included	Removing probable mapping errors with extremely high read depths
VCFtools	--not-chr	A chromosome or scaffold to be excluded	Removing a certain part of the genome (in this study a sex chromosome)

2.4.2 Comparability of sequencing batches

Measures were taken to prevent batch effects beforehand by choosing the same sequencing company and protocol for both batches, and by including five control samples that were sequenced in both batches to allow the detection of differences afterwards. As the controls originate from the same individual and the same DNA extract, any biological or preparation-related explanations for sequence differences can be ruled out.

To assess whether major sources of variation would be caused by batch effects, i.e. whether the sample individuals would group by sequencing batch rather than sample identity, a principal component analysis (PCA) was performed on the control samples. The replicate samples were assigned to two populations based on the sequencing batch (in the population map provided to the *populations* program). *Populations*

was run with the following settings: -p 2, -r 1, --min-mac 3, --max-obs-het 0.70, --write-random-snp, --blacklist (excluding the sex chromosome, see section 3.1.1, and all loci that contain SNPs with mean depth lower than 15 or higher than 130) (see Table 2 for the use of each setting). The PCA was carried out using the glPca function in the R package *adegenet* (v2.1.2; Jombart 2008, Jombart & Ahmed 2011). The results were plotted with *ggplot2* (Wickham 2011).

If a batch effect is not among the greatest sources of variability in the data, it cannot be detected using PCA only (e.g. Benito et al. 2004). To examine the presence of these potentially concealed effects, the default population genetic statistics calculated by Stacks were compared between the two batches. For the calculations, the data was filtered using the same set of filters as for the PCA, with the exception of using all SNPs within each locus. In order to rule out the possibility that coverage differences between batches were affecting genotype calls, the mean coverage among samples was normalized by downsampling to match the sample with lowest coverage (as in e.g. Regier et al. 2018). The initial coverage of each sample (BAM file) was obtained from Stacks after the alignment to reference genome, and the downsampling ratio was calculated by dividing the lowest coverage in the data with the coverage of the sample in question. The acquired ratio was used as the PROBABILITY parameter in the Picard DownsampleSam tool (v2.21.4; Broad Institute, https://broadinstitute.github.io/picard), with RANDOM_SEED = 1 and STRATEGY = Chained. After downsampling, the Stacks pipeline was followed as above and the population genetic statistics were calculated with corresponding filtering settings. The statistical significance of the difference in mean heterozygosity between batches was tested using a Welch two-sample t-test (e.g. Lu & Yuan 2010).

2.5 Population genetic analyses

2.5.1 Range edge population structure

The population structure along the spatial gradient in the Finnish range edge population was inferred using two complementary methods: principal component analysis (PCA), and a model-based Bayesian clustering approach with the program fineRADstructure (v0.3.2; Malinsky et al. 2018). The main purpose of the principal component analysis was to explore and visualize the structure in the data using unlinked SNPs. The fineRADstructure program utilizes haplotype data for inferring a coancestry matrix, a summary of nearest neighbor haplotype relationships in the data set. The program combines the coancestry matrix with a Markov chain Monte Carlo (MCMC) clustering algorithm (from the program fineSTRUCTURE; Lawson et al. 2012) into a tool for inferring population structure specifically from RAD-seq data. The use of haplotype

linkage information within each RAD locus provides a high resolution in comparison with methods that rely on unlinked SNPs.

In PCA, all 84 Finnish reed warbler individuals were included as a single population (in the population map provided to the *populations* program). *Populations* was run with the following settings: -p 1, -r 0.80, --min-mac 3, --max-obs-het 0.70, --write-random-snp, --blacklist (sex chromosome, loci that contain sites with lower mean depth than 15 and higher mean depth than 120) (see Table 2 for the use of each setting). The PCA was performed and plotted as in section 2.4.2.

For the fineRADstructure analysis, the same population map (including all 84 Finnish individuals as a single population) was used. *Populations* was run with the same settings as for the PCA, but this time all variable sites within each locus were retained, and the population filters -p and -r were applied haplotype-wise (-H) to reduce missing data within the haplotypes. The output file (populations.haplotypes.tsv) was converted into a fineRADpainter input file using a python script provided in the fineRADstructure package (Stacks2fineRAD.py, https://github.com/millanek/fineRADstructure). This step was run twice with two different filtering thresholds: -n 5 and -n 10, allowing a maximum of 5 or 10 SNPs per locus, accordingly. Increasing the allowed number of SNPs per locus can increase the resolution for detecting structure but also increase the risk of error, as too many SNPs at a locus may result from misassembled paralogs. The data was passed to the RADpainter, implemented in the fineRADstructure package. The MCMC clustering algorithm was run with 100 000 burn-in iterations, followed by 100 000 iterations sampled every 1000 iteration steps (-x 100000, -y 100000, -z 1000). The tree-building algorithm was run with 10 000 iterations of the algorithm to assess genetic relationships among clusters (-m T, -x 10000). Parameter values were selected according to the model developers' example. The results were visualized using the R scripts provided in the package (fineRADstructurePlot.R and FinestructureLibrary.R).

After running the analyses with all samples, three outlier individuals were removed (one individual from each detected outlier pair, see section 3.2) and the PCA and fineRADstructure analyses were re-run with this new population map (81 individuals), using the analysis-specific settings for *populations* and fineRADstructure as above.

2.5.2 Comparison of genetic diversity

Two areas of the same geographical size, one from the range edge (Finland) and the other from the range core (Czech Republic and Slovakia) were chosen for the genetic diversity comparison (Fig. 3). An equal number of individuals was used from both areas (n = 19 from edge and n = 19 from core) in order to

eliminate the effect of sample size on the diversity measures. As the range core samples were collected from three sampling sites, equal numbers of Finnish individuals were picked from three sites with corresponding spatial distances. Before running the diversity analysis, a fineRADstructure analysis was performed for the range core samples in order to confirm that the Czech and Slovakian individuals can be treated as a single, homogeneous population. No distinct genetic clustering was detected (analysis settings and fineRADstructure plot in Appendix A), and therefore the individuals were treated as one population in the diversity analysis.

Population genetic statistics for the two groups were calculated using the *populations* program in Stacks. The samples were assigned to two populations: range edge and range core. To measure genetic diversity, expected heterozygosity (H_e), nucleotide diversity (π), and the number of private alleles were measured for both populations. The chosen statistics rely on allele frequencies and counts, which makes them less susceptible to the bias from the batch effect than those relying on homozygosity or heterozygosity measures (see section 3.1.2). Expected heterozygosity, also known as gene diversity, describes the proportion of heterozygous genotypes expected under Hardy-Weinberg equilibrium, i.e. the probability that two randomly sampled allele copies from a population are different (Nei 1973). Nucleotide diversity is defined as the number of nucleotide differences per site between two randomly chosen sequences from a population (Nei & Li 1979), and its mean value is calculated of all pairwise comparisons. The number of private alleles is a measure of genetic distinctiveness, denoting the number of unique alleles in a population (term first used in Neel 1973). Stacks calculates the statistics using the following formulas:

$$H_e = \sum_{i \neq j} 2p_i q_j$$

Where *p* and *q* are the allele frequencies for the major and minor nucleotide, respectively.

$$\pi = 1 - \frac{\sum \binom{n_i}{2}}{\binom{n}{2}}$$

Where n_i is the count of allele i in a population, and n is the total count of alleles in a population at that locus (Nei & Li 1979, the presented formula used by Stacks in Catchen et al. 2013).

Populations was run with the following settings: -p 2, -r 1, -H, --min-mac 3, --max-obs-het 0.70, --blacklist (sex chromosome, loci that contain sites with lower mean depth than 15 and higher mean depth than 130). The setting -r 1 specifies that the analyzed loci must be present in all individuals in both populations, which prevents missing data from affecting the diversity estimates.

3. RESULTS

3.1 Data quality and comparability

3.1.1 Sex chromosome detection

After initial filtering and data plotting, it became apparent that the samples were split into two distinct groups that did not correspond to spatial sampling locations. This can be seen in the PCA of the Finnish samples (Fig. 4). The division was confirmed to be caused by the sex of the individuals by comparing each individual in the PCA to its assigned sex in field sampling. The division matched the assigned sex of the birds in all but three individuals, which is very probably explained by sexing mistakes. The reed warbler is a monomorphic species, meaning that the male and female look similar. The sex was determined in the field by direct observations on singing, which is characteristic for males, and the presence of a brood patch (a patch of featherless skin on the belly for egg incubation). While typically only females develop a brood patch, it may not be present in all individuals or its presence may not always be clear. Also males can express a partial brood patch and do not always sing before catching (Katja Rönkä, personal communication 2020), making in-field sexing difficult in this species.

As the used reference genome was assembled to scaffold- but not chromosome level, the scaffolds that correspond to the sex division needed to be identified. From the PCA of the Finnish population, allele loadings were retained for the principal component that captured the sex-based variation (PC1, Fig. 4) to identify the top loading markers and the scaffolds they are located in. The majority of the highest loading markers were found to be located on scaffold 23 in the reference genome, suggesting that it is the sex chromosome (likely a fusion of the Z and W chromosomes; Pasi Rastas, personal communication 2020) responsible for the most genetic variation between sexes. Excluding this scaffold removed the sex division pattern from the data, and therefore it was excluded from all further analyses to prevent it from obscuring the population genetic patterns of interest (e.g. Benestan et al. 2017).

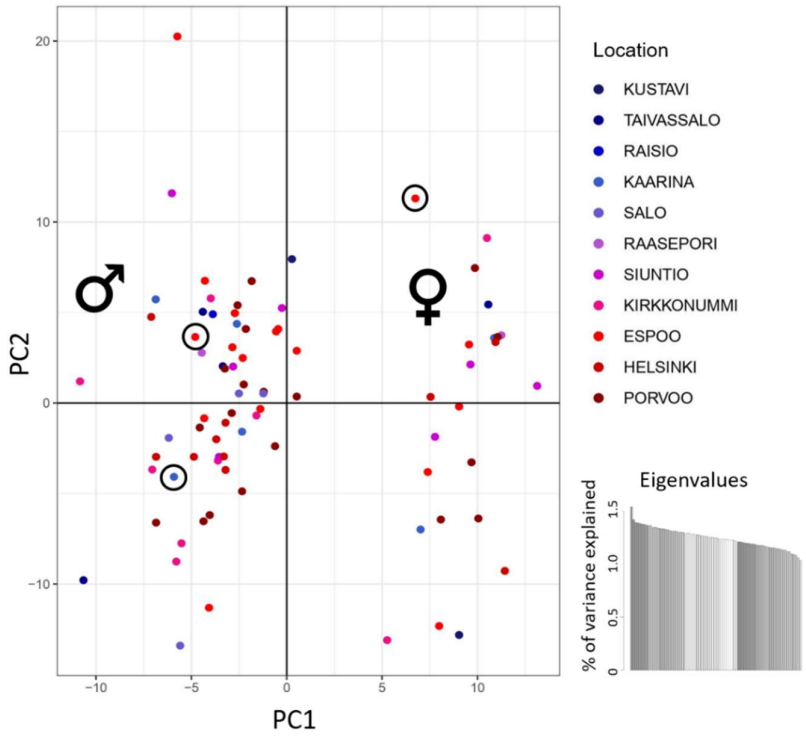


Figure 4. PCA of Finnish samples before removing the sex chromosome, showing a division by sex on the PC1 axis. Each point represents one sample individual, and the point is coloured according to the sampling municipality. The circled individuals were assigned a different sex during field sampling. Eigenvalues of each principal component are presented as the percentage of explained variance. Filtering settings for the *populations* program: -p 1, -r 0.80, --min-mac 3, --max-obs-het 0.70. Three outlier individuals (see section 3.2) were excluded from the analysis.

3.1.2 Batch effect detection

The data set for the PCA of the control samples contained 11 532 SNPs (1 SNP / locus). The PCA (Fig. 5 a) shows that samples cluster well by identity and not by batch, which suggests the comparability of the two sequencing batches in population structure analyses. Despite PCA showing no batch effect, the comparison of population genetic statistics gave differing results across the sequencing batches. There was an especially clear difference in the mean observed heterozygosity (Batch 1: 0.269, Batch 2: 0.142) and the statistics derived from it (e.g. F_{IS}, Batch 1: 0.272, Batch 2: 0.620 (Wright 1931; Catchen et al. 2013)).

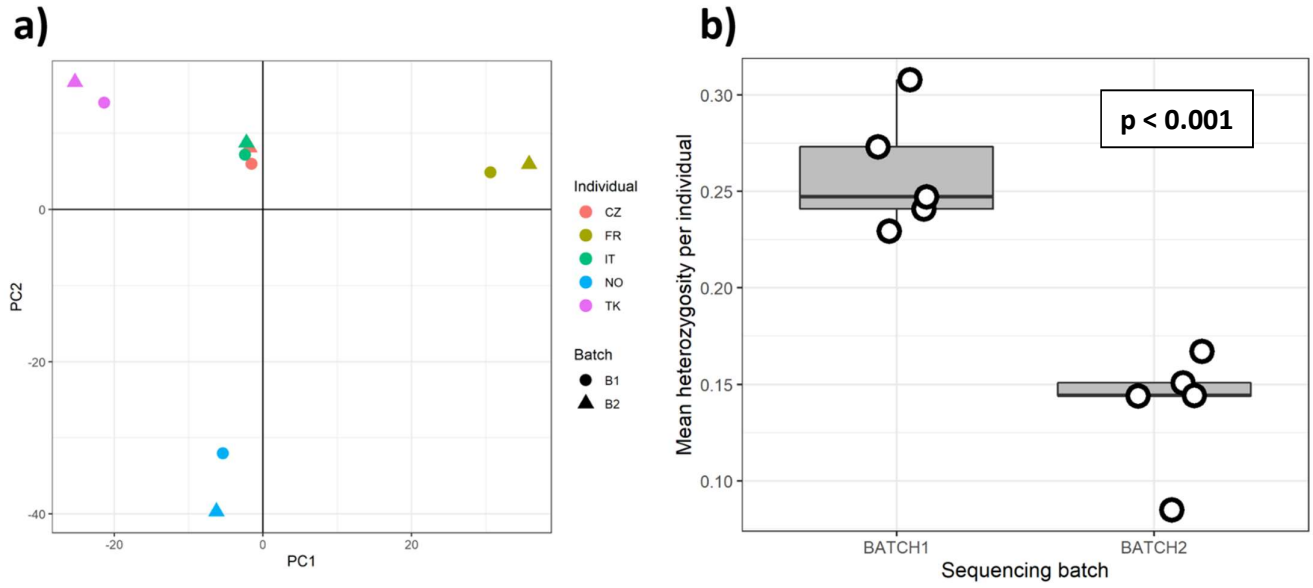


Figure 5. a) PCA of the five control individuals sequenced in both batches (two first axes are shown). Point colour indicates the identity of the control individual, and point shape indicates the batch that it was sequenced in (round = Batch 1, triangle = Batch 2). The names of the control individuals are abbreviations of their country of origin (CZ = Czech Republic, FR = France, IT = Italy, NO = Norway, TK = Turkey). **b)** Box plot of the mean observed heterozygosity per control individual across the two sequencing batches, after normalizing the coverage. The lower and upper hinges of the box correspond to the 25th and 75th percentiles, the thicker horizontal black line within each box is the median, and the whiskers extend to the largest and lowest value that are at most 1.5 * IQR (interquartile range) from the hinge. Each circle represents the mean heterozygosity value for a sample.

Even though both sequencing batches had approximately similar numbers of reads (Batch 1: 7.1 million reads/sample, Batch 2: 6.5 million reads/sample) the mean coverage (the number of unique reads per nucleotide) of the control samples in Batch 1 was almost twice as high as in Batch 2 even after filtering (100.6x and 53.0x, respectively). However, normalizing the coverage did not remove or decrease the difference in heterozygosity between the batches (Fig. 5 b). The mean individual heterozygosity in Batch 1 was 0.260 (SD = 0.194), in Batch 2 it was 0.138 (SD = 0.164). A Welch two-sample t-test showed that the difference was statistically significant, $t(8.00) = 6.14$, $p < 0.001$. The full set of statistic comparisons can be seen in Appendix B.

A probable cause for the observed batch effect are PCR duplicates in Batch 2, resulting in false homozygous calls. This was also suspected in a quality report produced by the sequencing company. Unfortunately, these duplicates cannot be separated and removed from data that is sequenced using the original RAD-seq protocol (Andrews et al. 2016). As filtering or downsampling procedures did not remove the batch effect, it will affect the comparability between samples across the two sequencing batches. Especially analyses that

rely on homozygosity or heterozygosity estimates cannot reliably be used to report biological differences between populations. How this affects the results of this study is further addressed in section 4.3.

3.2 No spatial population structure at range edge

A principal component analysis (PCA) and a fineRADstructure analysis (a model-based Bayesian clustering approach) were performed to evaluate population structure in the recently established range edge population of reed warblers in Finland. The analyses were based on sequence data from 84 individuals, collected along a spatial gradient (west–east) along the southern coast of Finland. No spatial structuring was detected in either of the analyses; the population was homogeneous across the sampled area. However, both analyses identified three pairs of individuals that were genetically closer to each other than to the rest of the samples.

After applying the filtering criteria, the data set for the PCA contained 37 929 unlinked SNPs (1 SNP / locus; missingness per individual 0.8 – 23.0 %). The PCA (Fig. 6 a) shows that both the western (blue) and eastern (red) samples are clustered in a single group, except for three pairs of individuals that are each clustered separately from the rest of the samples. These outlier pairs can also be seen in the eigenvalue plot: the eigenvalues of the three first principal components stand out, explaining the largest shares of the variance in the data. Within each pair, the two individuals are from different sampling locations.

The fineRADstructure analysis was based on a set of 120 194 SNPs within 49 460 loci (missingness per individual 0.9 – 22.8 %), the SNPs within each locus forming a haplotype. The fineRADstructure program processed ca 20 700 loci with the threshold of 5 SNPs per locus, and ca 41 600 loci with 10 SNPs per locus. Increasing this threshold did not change the outcome of the analysis: the population tree and coancestry matrix using 5 SNPs per locus are shown in Fig. 6 b, and the corresponding plots with 10 SNPs per locus can be found in Appendix C. The results of the fineRADstructure analysis reflect the results of the PCA, showing no spatial clustering of the samples but identifying the same, three pairs of individuals with higher pairwise coancestry than the other samples. In the tree, only these three pairs have greater than 95 % branch support. The posterior probabilities of most branches are zero or close to zero. A darker yellow vertical band can be seen near the centre of the heat map, consisting of a group of individuals with slightly higher estimated coancestry with all samples. These were found to be the individuals with least missing data, indicating that the apparent, although not well-supported grouping only reflects a technical feature of the samples.

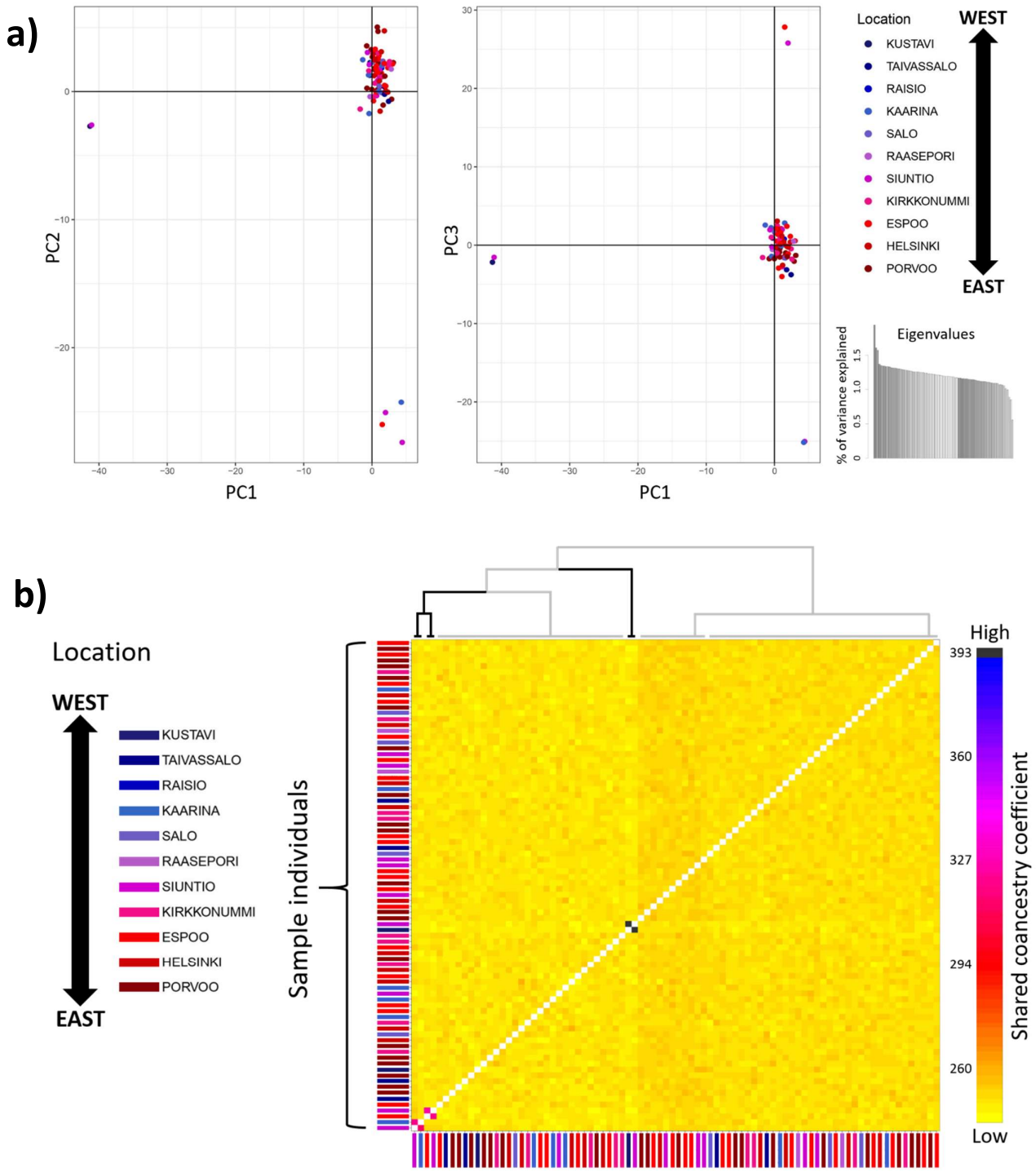


Figure 6. There was no spatial clustering of *Acrocephalus scirpaceus* samples along a west–east gradient at the northern range edge of the species (southern coast of Finland). **a)** Principal component analysis (PCA) of the genomic data (37 929 SNPs). The scatterplots show factor scores of individuals in first and second (left) and first and third principal components (right). The percentage of explained variation by each PC is shown in the eigenvalue plot. The colour of each individual shows its sampling location. The locations are coloured according to their position on the spatial gradient: blue in the west, purple in the middle and red in the east. **b)** FineRADstructure analysis of the data (20 700 RAD loci). In the population tree, branches with higher than 95 % posterior probability are shown in black, grey branches indicate support below this limit.

In the heat map, each small square indicates a pairwise coancestry value between individuals (number of loci for which the two individuals are closest in terms of genetic distance): lowest coancestry estimates in the data are represented by yellow, intermediate by red, and highest by blue and black. Sample individuals are plotted as small bars on the left side and under the heat map: the colour of each bar represents the sampling location of the individual, shown on the left.

After removing one individual from each of the three outlier pairs, the data set for the PCA contained 38 960 SNPs (1 SNP / locus; missingness per individual 0.8 – 23.3 %). All 81 individuals, including the remaining individuals from the previously detected outlier pairs, form a single cluster in the PCA (Fig. 7 a). This shows that the outliers are not genetically different to the rest of the population, but stand out because of high within-pair similarity. Within the single cluster in the PCA, samples from eastern and western sampling sites seem to be homogeneously mixed. No spatial clustering can be seen along the first three principal component axes, indicating that the sample location on the west – east gradient does not account for the observed variation in the data. None of the principal components explains markedly more variation than the others, which can be seen in the eigenvalue plot (Fig. 7 a).

The corresponding haplotype data set for fineRADstructure contained 122 884 SNPs within 51 003 loci (missingness per individual 0.9 – 23.3 %). Of these, ca 21 300 loci were processed by fineRADstructure with the threshold of max 5 SNPs per locus, and ca 42 900 loci with max 10 SNPs per locus. Increasing the SNP threshold did not change the results except for identifying one more pair of individuals with slightly higher relatedness than the rest (Appendix C). The results using 5 SNPs per locus (Fig. 7 b) are very similar to the results from the PCA (Fig. 7 a): the estimated coancestry values between individuals fall now within a narrower range, and no genetic clusters can be identified. None of the population tree branches are well supported (> 95 % posterior probability), and the cluster with the highest support (66 %; seen in the plot as a vertical band with more purple) again corresponds to the individuals with least missing data, not a biologically relevant group.

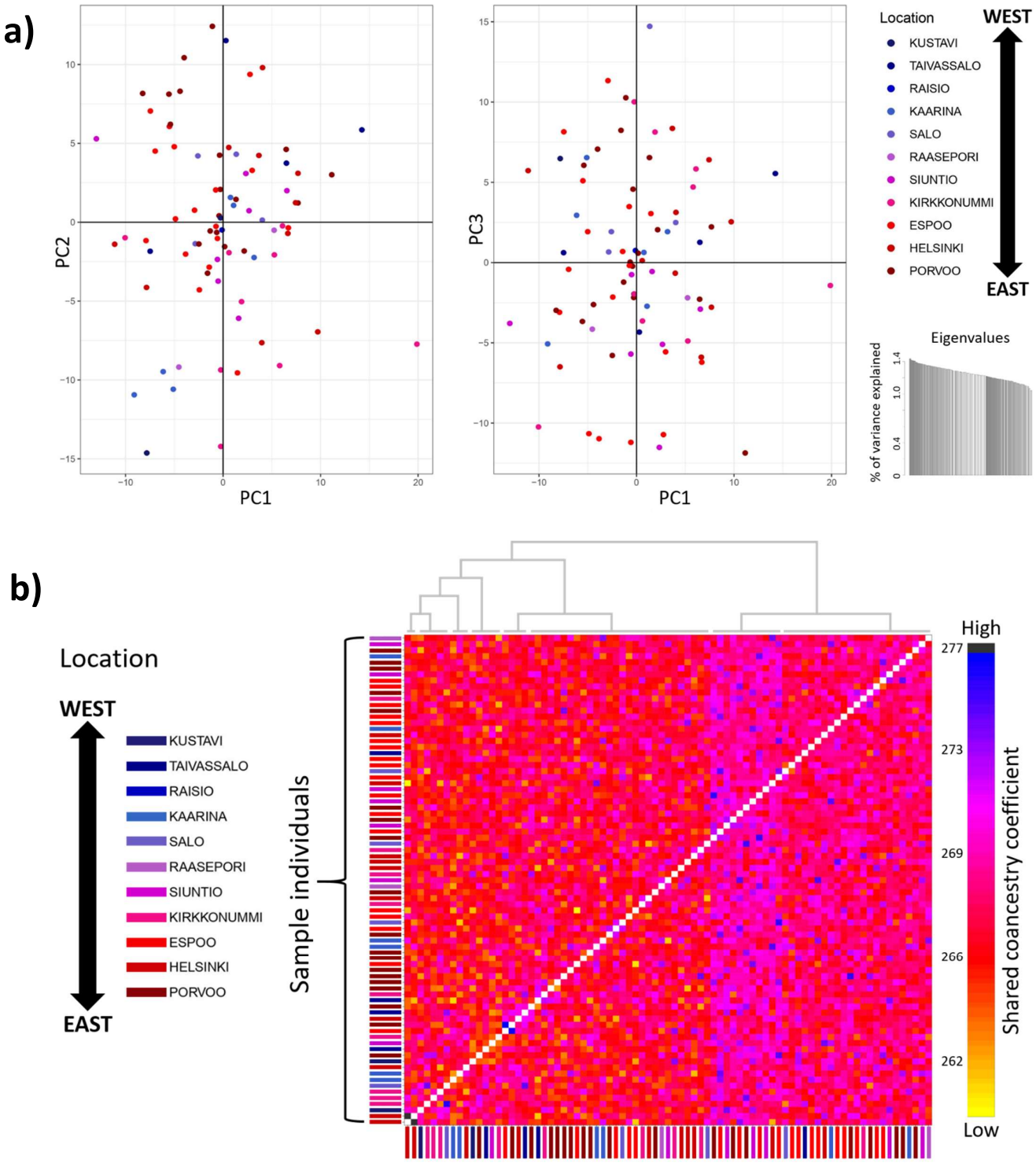


Figure 7. The results of the population structure analyses after removing one individual from each outlier pair. No spatial clustering the samples can be detected. **a)** Principal component analysis (PCA) of the genomic data (38 960 SNPs). The scatterplots show factor scores of individuals in first and second (left) and first and third principal components (right). The percentage of explained variation by each PC is shown in the eigenvalue plot. The colour of each individual shows its sampling location. The locations are coloured according to their position on the spatial gradient: blue in the west, purple in the middle and red in the east. **b)** Output of fineRADstructure analysis (21 300 RAD loci). In the population tree, all posterior

probability values are below 95 % (grey branches), i.e. no branches are well-supported. The heat map indicates pairwise coancestry between individuals (number of loci for which the two individuals are closest in terms of genetic distance): lowest coancestry estimates in the data are represented by yellow, intermediate by red, and highest by blue and black. Sample individuals are illustrated as small bars on the left side and under the heat map: the colour of each bar represents the sampling location of the individual, shown on the left.

3.3 Similar levels of genetic diversity at the range edge and range core

The genetic diversity measures at the range edge (Finland) and range core (Czech Republic + Slovakia) were calculated using a data set of 10 186 SNPs within 7448 loci (38 samples, 19 indv. from edge and 19 indv. from core). Perhaps surprisingly, the comparison shows no signs of reduced genetic diversity during the species' northward range expansion in Europe (Table 3). Of the 10 186 sites that are variable in the whole data set, over 90 % are also polymorphic in each of the two populations (i.e., have two or more alleles within the specified population). The fraction of polymorphic sites is only 0.57 % higher in the range core population. The number of private alleles is slightly higher in the range core samples than in the range edge samples. The mean values of expected heterozygosity (H_e) and nucleotide diversity (π) fall within the same range at both range edge and range core, being even slightly higher in the range edge population ($H_e = 0.198$, $\sigma^2 = 0.020$; $\pi = 0.204$, $\sigma^2 = 0.021$) than in the range core population ($H_e = 0.195$, $\sigma^2 = 0.019$; $\pi = 0.200$, $\sigma^2 = 0.021$).

Table 3. Population genomic parameters reflecting the genetic diversity of reed warblers at the range edge and range core areas. FI = Finland, CZ = Czech Republic, SK = Slovakia, N = number of sampled individuals, H_e = expected heterozygosity under Hardy-Weinberg equilibrium, π = an estimate of nucleotide diversity, SE = standard error.

Population	N	Variant sites	% Polymorphic sites	Private alleles	$H_e \pm$ SE	$\pi \pm$ SE
Range edge (FI)	19	10 186	91.95	726	0.198 ± 0.001	0.204 ± 0.001
Range core (CZ + SK)	19	10 186	92.52	820	0.195 ± 0.001	0.200 ± 0.001

However, the detected batch effect (section 3.1.2) might inflate the estimates of H_e and π at the range edge. In the comparison of the five control samples from the two sequencing batches, the estimates for H_e and π were both approximately 4 % higher in the batch with the higher observed homozygosity (the batch with the Finnish samples) than in the other batch (the batch with the Central European samples) (Appendix B). If the relative error is in this range, the estimate of genetic diversity at the range edge would still be close to the diversity at the range core.

4. DISCUSSION

Theoretical studies generally predict range expansions to cause genetic changes in the expanding populations, including population structure and reduced genetic diversity along the expansion axis (reviewed in Excoffier et al. 2009). These effects mainly result from recurring founder events and strong genetic drift at the range front, and therefore the strength of gene flow is expected to affect the degree of the genetic consequences (Ray et al. 2003, Excoffier 2004; Excoffier et al. 2009). Despite of a large body of theoretical research, empirical studies on the genetic consequences of environment-driven range expansions, especially in highly mobile species, are scarce. This study provides an example from a recently founded population of a migratory passerine, the Eurasian reed warbler (*Acrocephalus scirpaceus*), which appears to have largely avoided the predicted genetic changes of range expansion. The use of a fine-resolution RAD-seq data set showed a lack of population structure and high levels of retained genetic diversity at the northern range edge of the species. In the following sections, I discuss the main findings of this study and how they relate to our current knowledge of range expansions. Additionally, I discuss the detected batch effect and the need for routinely including controls in genetic studies.

4.1 Dispersal propensity and habitat connectivity may underlie the lack of population structure at the range edge

Both simulation studies (e.g. Rendine et al. 1986, Fix 1997, Currat & Excoffier 2005) and empirical studies (e.g. Short & Petren 2011, Swaegers et al. 2013) have shown that range expansions can create allele frequency gradients and therefore population structure. These gradients can arise as a result of colonisations with admixture with local individuals, repeated founder effects in the absence of admixture, or kin-structured colonisations, where the colonizing individuals are closely related (reviewed in Excoffier et

al. 2009). While the presence of genetic structure after a range expansion seems to be more common in species with lower mobility, it has also been reported from highly mobile species, such as monarch butterflies (Pierce et al. 2014) and coyotes (Heppenheimer et al. 2018). In this study, no genetic structuring could be detected in the range edge population of the reed warbler, analysed along a spatial gradient (ca 250 km) representing the assumed southwestern colonization route of the species (Leivo 1937, Wikström 1945). This is contrary to the primary hypothesis, which predicted that the high-resolution genetic structure would reflect the historical advancement of the expansion, based on theoretical expectations and the relatively high site fidelity of the species. Instead, the lack of spatial structure is in line with the alternative hypothesis that higher levels of gene flow or more frequent LDD events may have prevented or broken down the patterns of genetic structure.

Studies using spatial gradients of similar length have detected genetic structure after equally recent range expansions in other species, including a more sedentary bird, the hazel grouse, in the French Alps (Rózsa et al. 2016). The lack of genetic structure in the Finnish reed warbler population is likely to be indicative of relatively high gene flow between the sampling sites, and one plausible reason for this is the dispersal propensity of the species. Migratory species and species living in wet habitats tend to disperse further than resident species or species living in dry habitats (e.g. Paradis et al. 1998). Despite being philopatric, reed warblers have been reported to make occasional long-distance dispersal (LDD) events up to 300 km (Paradis et al. 1998). For comparison, the longest documented dispersal distances for the hazel grouses in France have been 25 km (Montadert & Léonard 2006). LDD events have been hypothesized to be the cause for low differentiation and weak spatial genetic structure during an invasion of European starlings (Berthouly-Salazar et al. 2013) and a range expansion of melodious warblers (Engler et al. 2015), both of which are migratory passerines. The relatively frequent natal and breeding dispersal events of the reed warbler have also been suggested to be the reason for low overall population differentiation across the species' range (Paradis et al. 1998, Procházka et al. 2011). The role of LDD events in attenuating genetic structure during colonization has also been pointed out by theoretical studies (e.g. Bialozyt et al. 2006). Additionally, it is possible that dispersal abilities are under positive selection during range expansions: selection has been shown to favour individuals with even greater dispersal abilities at the range margins (Travis & Dytham 2002, Berthouly-Salazar et al. 2012). Whether this has occurred during the reed warbler range expansion is not known.

In addition to dispersal, environmental homogeneity presents another possible reason for the homogeneous population structure in the sampled area. The environment along the southern coast of Finland is relatively continuous in terms of climatic conditions and the presence of suitable reed bed habitat, which might have promoted successful dispersal events and gene flow between sample sites (e.g. Sexton et al. 2014). What might also suggest high connectivity between the sampling sites is the detection

of three pairs of individuals with high within-pair similarity. Interestingly, none of the pairs were sampled within the same municipality: the greatest geographic distance between the individuals in a pair was ca 150 km (sampling municipalities Kustavi and Siuntio). The grouping of each pair resulted from high pairwise genetic similarity and not from genetic distance to other individuals, as removing one individual from each pair made the remaining samples cluster together with the rest. This is likely to indicate high relatedness between the individuals in each pair, which would mean that closely related individuals could disperse to the opposite ends of the sampling gradient in just a few generations, which would also be possible based on the documented dispersal distances in other regions (Paradis et al. 1998). However, as no further relatedness analyses were performed as a part of this study, the possibility of sample contamination as a cause for the genetic similarity cannot completely be ruled out.

Thus far only a few published studies have explored the genetic patterns of environment-driven range expansions in avian species. More examples come from introduced non-native expansions (e.g. house sparrows; Schrey et al. 2011, European starlings; Berthouly-Salazar et al. 2013), but these are not fully comparable to expansions at native range borders, as introductions are often characterized by a very small group of founder individuals, and limited to no gene flow from other parts of the range. In addition to hazel grouses (Rózsa et al. 2016) and melodious warblers (Engler et al. 2015), genetic structure after recent or ongoing range expansion has been studied at least in light-vented bulbuls (Song et al. 2013), European bee-eaters (Ramos et al. 2016), and common cranes (Haase et al. 2019). In light-vented bulbuls, range front populations exhibited less structure than range core populations, which is contrary to the theoretical prediction that spatial expansions would induce structuring in the new populations. Slight to moderate range-wide structuring was detected in crane and bee-eater populations, but these studies did not specifically analyse spatial structure at range edges. In the context of this study, I also analysed the European-wide population structure of reed warblers from all available data (Appendix D). The patterns seem to resemble those of the light-vented bulbuls: despite low levels of differentiation (Procházka et al. 2011), range core populations show relatively well-defined spatial structure in many areas. The range edge populations (Finland and Norway), however, form a single cluster with no apparent within-group differentiation, even though there is differentiation between this cluster and the range core populations. These findings seem to support the results of the few previous studies, suggesting that a lack of population structure at range edges might be a generalizable pattern in species with sufficiently high dispersal rates, possibly accompanied with LDD. Additionally, all the mentioned studies on avian range expansions have used DNA microsatellite or mitochondrial markers. This study shows that even with increasing the resolution to tens of thousands of SNPs or haplotypes, no spatial structure could be detected in the sampled range edge population of this mobile species.

4.2 Retained genetic diversity in highly mobile species: exception or expectation?

Loss of genetic diversity is another theoretically expected consequence of restricted dispersal during range expansions. Experiencing multiple founder events or allelic surfing can reduce genetic variation in marginal populations due to increased drift (Austerlitz et al. 1997, Klopfstein et al. 2006; Excoffier et al. 2009). Reduced genetic diversity along natural expansion axes has indeed been observed in a wide range of taxa, from plants (e.g. Pujol & Pannell 2008, Keller et al. 2010) to vertebrates (e.g. White et al. 2013, Garcia-Elfring et al. 2017). Gene flow and therefore dispersal greatly impact the expected diversity patterns, and frequent LDD events can preserve genetic diversity during a range expansion. In this study, almost no signs of reduced genetic diversity were detected in the recently founded range edge population of reed warblers, contrary to the primary hypothesis. The number of private alleles was slightly higher in the range core population (Czech Republic and Slovakia) than at the range edge (Finland), but the estimated values for expected heterozygosity (H_e) and nucleotide diversity (π) were very similar between edge and the core populations. This indicates that reed warblers have managed to retain most of the genetic variability during the species' expansion through Northern Europe, suggesting high levels of gene flow not only at small spatial scales (such as the Finnish coast), but also over wider geographical range. Due to a detected batch effect that reduced heterozygous genotype calls in the range edge population, the observed heterozygosity and inbreeding coefficients could not be reliably compared between range edge and core (see section 4.3).

An important consideration about the reliability of any genetic diversity comparison is the representativeness of the used samples. How well the chosen samples capture the actual diversity in the focal populations depends both on sampling design and sample sizes. In this study, a single area was chosen from the available data to represent the genetic diversity at the range core. The choice was based both on the suitable size and location of the area, and on results from a previous study that reported similar levels of mitochondrial diversity across the core areas of the established reed warbler range (Arbabi et al. 2014). An equal number of individuals from the range edge and the range core, and only genomic sites that were present in both populations were used to eliminate biases caused by differences in sample size and the amount of missing data. The sample size, 19 individuals and > 10 000 SNPs should be sufficient to accurately estimate the genetic diversity within each group: Nazareno and colleagues (2017) found that when using > 1000 SNPs (RAD-seq), increasing the sample size above eight individuals had little impact on the diversity estimates in a tropical plant species. This is likely to be applicable for passerines as well, as the suitable sample size for microsatellite studies (using < 10 loci) has been found to be 20–30 individuals (song sparrows; Pruett & Winker 2008).

True genetic diversity in RAD-seq data can be underestimated due to allelic dropout: the most variable genomic areas are also most likely to have mutations at the restriction enzyme cut sites, and therefore they are more likely to be missed during sequencing (Arnold et al. 2013). However, the bias has been shown to be of minor importance in systems with low polymorphism (< 2 %) (Cariou et al. 2016), which is the case in both the range edge and core populations of reed warblers in this study (polymorphic loci of all sequenced loci 1.34 % and 1.35 %, respectively). Thus, allelic dropout is not likely to be an important source of bias in the genetic diversity estimates in this study. Considering the detection of rare alleles, in this study only alleles that were present as at least three copies across all samples were included in the analyses. While removing potential genotyping errors, some of the rarest alleles may also be excluded from the diversity measures.

As with population structure, published literature contains only few examples of how recent range expansions have affected genetic diversity in avian species. Engler and colleagues (2015) reported no changes in genetic diversity at the expanding range margin of melodious warblers. Song and colleagues (2013) found that different range front populations of light-vented bulbuls harboured both low and high levels of genetic diversity, therefore showing no statistically significant relationship between lowered diversity and range expansions. Introduced avian species have in many cases shown patterns of reduced diversity compared to source populations (e.g. Cabe 1998, Hawley et al. 2006), but these populations are deprived of the gene flow that could occur within native ranges. The preliminary results of this study are in line with the existing examples on environment-driven shifts in migratory passerines. They suggest that high mobility and the tendency for sufficiently frequent LDD can allow species to escape one of the most severe genetic costs of a range expansion: the depletion of genetic diversity. As genetic diversity is tightly linked with adaptive potential in a new environment (e.g. Barrett & Schluter 2008), the ability to preserve diversity during a range expansion might be a factor that reduces extinction risk in rapidly changing or fragmenting environments. More studies are needed for testing how commonly and in what conditions highly mobile species are able to retain high levels of genetic variation during range expansions.

4.3 Batch effect and the reliability of the results

In addition to giving insights into range expansions of mobile species, this study provides an important reminder of the need to account for potential biases in the data. One very common and potentially very problematic complication in high-throughput sequencing is the presence of batch effects (Leek et al. 2010). Batch effects can arise from even minor technical differences in the handling or sequencing of the samples. These differences introduce a source of variation between batches that is unrelated to real biological

variation in the data. If not accounted for, these differences may be misinterpreted as meaningful biological signals, even leading to the publication of erroneous conclusions (e.g. Sebastiani et al. 2011; reviewed in Leek et al. 2010). A positive aspect of this issue is that unlike some other genetic techniques, high-throughput technologies provide enough data for detecting batch effects, and different methods for removing them have been developed. Unfortunately, these methods are not perfect: they may not fully remove the batch effect or they may simultaneously remove biologically relevant variation (Goh et al. 2017). Therefore, treating batch effects afterwards cannot replace good experimental design. However, as it is probably impossible to avoid batch effects completely, a good experimental design (e.g. inclusion of technical replicates or randomization of samples between batches; O'Leary et al. 2018) is also the only way to detect them and deal with them later on.

In this study, the inclusion of five replicate samples (i.e. controls) across two sequencing batches enabled the detection of a batch effect. Despite using the same sample preparation protocol and sequencing company in both batches, there was a significant excess of homozygous genotype calls in the batch that contained the Finnish range edge samples. Without the comparison of control samples, there is a possibility that this technical artefact could have been interpreted as a biological difference between range edge and range core populations: a decrease in heterozygosity along the expansion axis is also a possible consequence of a range expansion (Austerlitz et al. 1997). It also has to be noted that this batch effect was not apparent in the PCA of the control samples, which is a commonly recommended method for batch effect detection (e.g. O'Leary et al. 2018). Instead, here the comparison of population genetic statistics of the replicate samples revealed the difference.

The difference in heterozygosity between batches was suspected to be caused by PCR duplicates. These seem to have excessively amplified one but not the other allele in some of the true heterozygous genotypes, resulting in false homozygous genotype calls. However, these excess amplifications seem to have occurred more or less randomly, as the statistics derived from allele frequencies are much less affected than statistics derived from genotypes (see section 3.1.2 and Appendix B). The correct grouping of control samples in the PCA (Fig. 5 a), and the grouping of Finnish and Norwegian samples (sequenced in different batches) in the European-wide structure analyses (Appendix D: Fig 12) seem to suggest that population structure can quite reliably be inferred from the data. It is however possible that the falsely duplicated alleles somewhat reduce the precision of the population structure analyses in the Finnish samples. In the genetic diversity analyses, only statistics derived from individual allele frequencies and counts are reported. In the replicate samples, the estimates for these statistics (H_e and π) were nevertheless approximately 4 % higher in the batch containing the range edge samples. Therefore the diversity results are considered preliminary, although they likely are suggestive of the true diversity. No

systematic occurrence of false heterozygotes (i.e. artifactual SNPs) was detected, suggesting that polymorphism and private alleles are not overestimated at the range edge.

To conclude, the inclusion of replicate samples lead to the discovery of a significant batch effect, which may have biased the results had it gone undetected. Population structure analyses seem to be robust across batches, and only the least affected genetic diversity statistics are reported as preliminary results. The analyses will be repeated with re-sequenced data in the near future. While technical replicates and the randomization of samples across batches are sometimes overlooked in genetic studies (Leek et al. 2010), they should be a standard practice for detecting and managing batch effects in order to avoid false or misleading conclusions.

4.4 Range expansion success in a rapidly changing world

The rapid and pervasive anthropogenic environmental change is causing range shifts across different taxa (Parmesan & Yohe 2003, Chen et al. 2011). It has therefore become essential to understand and predict how different species are able to respond to the changing conditions, and what traits of the species and the environment facilitate successful range shifts. In order to shift its range, a species must both be able to reach a novel environment by dispersal and establish a population there. It is hardly an exaggeration to consider the northward range expansion of the reed warbler thus far successful: during only ca 200 years, the species has been able to establish large populations across Denmark and the southern and central parts of Fennoscandia (Wikström 1945, Løppenthin 1967: cited in Avilés et al. 2006, BirdLife International 2021). The results of this study suggest that the expanding population has been able to retain high levels of genetic diversity, which would be in line with the theoretical predictions of diversity as a contributor in colonization success (e.g. Crawford & Whitney 2010, Wennersten & Forsman 2012, Szűcs et al. 2014). On the other hand, strong gene flow from range core to range edge populations has been suggested to also have contradicting effects by swamping local adaptation (e.g. Bridle & Vines 2007, but see Kottler et al. 2021, in press, for a review of empirical evidence). How or whether genetic variability has facilitated the expansion of reed warblers remains to be solved.

As genetic variability increases the adaptive potential of a population, loss of variation during a range shift can hinder adaptation to the novel conditions at the range edge, slowing down the expansion or even leading to the collapse of marginal populations (Pujol & Pannell 2008, Peischl et al. 2015, Szűcs et al. 2017, Polechová 2018). The need for adaptation and therefore genetic variation may be greatest when there is a steep environmental gradient between the previously occupied range and the new area, i.e. more differences between the characteristics of the sites (e.g. Polechová 2018). In Northern Europe, climatic

conditions have changed (IPCC 2014) and the presence of reed bed habitat has increased (Altartouri et al. 2014), making the environmental gradient less steep from the perspective of the reed warblers. Although the environment is certain to differ in some ways from the previously occupied range, it seems possible that the recent expansion of reed warblers may have mostly been driven by spatial tracking of the existing niche instead of niche evolution.

However, genetic diversity may promote range shifts also in other ways than increasing adaptive potential and the likelihood of possessing pre-adapted genotypes in novel environmental conditions. The adaptive potential can also be related to expansion-facilitating traits, such as higher dispersal, increased population growth rates or aggressive behaviour (e.g. Duckworth 2008, Szűcs et al. 2017). By increasing genetic diversity, gene flow during a range expansion can also attenuate the costs of founder effects and inbreeding, including the accumulation of homozygous, deleterious alleles (e.g. González-Martínez et al. 2017). Overall, the benefits of at least moderate gene flow during colonization seem to be apparent. Yet, genetic polymorphism is not the only factor than can cause phenotypic diversity. Developmental plasticity, randomized phenotype switching (i.e. bet-hedging), and behavioural plasticity (e.g. learning) can also create variation in traits and behaviour of individuals, and have all been associated with increased range expansion success (Sutter & Kawecki 2009, Tuomainen & Candolin 2011, Wennersten & Forsman 2012). Also reed warblers have been shown to modify their behaviour based on social information and in response to rapid changes in brood parasitism risk, supporting behavioural plasticity in the species (e.g. Davies & Welbergen, Thorogood & Davies 2013). Plastic responses to changing environmental conditions may be especially important for less mobile and sessile species (Wennersten & Forsman 2012), species with long generation times (e.g. Refsnider & Janzen 2012), and species inhabiting heterogeneous environments (Sutter & Kawecki 2009), allowing fast phenotypic changes and driving genetically-based evolutionary changes in the long run.

There is great variation in species' responses to ongoing environmental change, including range shifts. Meta-analyses have found that using only species' traits generally has low explanatory power for predicting the variation in range shifts, even though the traits (e.g. greater dispersal ability) and observed shifts would show statistically significant relationships (Angert et al. 2011, MacLean & Beissinger 2017). Therefore, it has been proposed that an ideal framework for predicting variation in range shifts should include the effects of niche tracking through space or time, plasticity or acclimation, evolution, and a standardized set of species' traits (MacLean & Beissinger 2017), thus containing also the genetic aspect of range shifts. Also other research syntheses have recently emphasized the importance of genetics in predicting ecological and evolutionary responses to environmental change (e.g. Harrisson et al. 2014, Bay et al. 2017). This study provides an example of the genetic effects of range expansion in a highly mobile passerine bird species, presenting results that are likely to be indicative of maintained evolutionary potential in the range edge

population. When combined with further studies on adaptation and the other factors of the suggested framework, the findings of this study may provide useful insights for understanding and predicting range expansions. Additionally, a population genetic comprehension will be valuable when studying other aspects of reed warbler ecology and evolution, such as the roles of evolutionary adaptation and behavioural flexibility in shaping responses to novel conditions and species interactions at the range edge.

ACKNOWLEDGEMENTS

I am grateful to my supervisors for guidance and support: Katja Rönkä for being an amazing mentor throughout this project and introducing me to the study system, Rose Thorogood for insightful comments and work behind the scenes to make all the research possible, and Perttu Seppä for valuable advice on population genetics. I want to thank the personnel of the Molecular Ecology and Systematics (MES) laboratory for the help and flexibility with getting everything finished before the lockdown; Biodata Analysis Unit and Pasi Rastas for advice on data processing; Helsinki Institute of Life Science (HiLIFE) for supporting this work financially by granting a research trainee scholarship; Fabrice Eroukhmanoff, Camilla Lo Cascio Sætre, Petr Procházka and Bård Stokke for collecting and providing samples from locations across the European range of reed warblers; and CSC – IT Center for Science, Finland, for computational resources. Finally, many thanks go to friends and family.

REFERENCES

Alagador, D., Cerdeira, J. O., & Araújo, M. B. 2016: Climate change, species range shifts and dispersal corridors: an evaluation of spatial conservation models. — *Methods in Ecology and Evolution* 7: 853–866.

Altartouri, A., Nurminen, L., & Jolma, A. 2014: Modeling the role of the close-range effect and environmental variables in the occurrence and spread of *Phragmites australis* in four sites on the Finnish coast of the Gulf of Finland and the Archipelago Sea. — *Ecology and evolution* 4: 987–1005.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. 2016: Harnessing the power of RADseq for ecological and evolutionary genomics. — *Nature Reviews Genetics* 17: 81.

Angert, A. L., Crozier, L. G., Rissler, L. J., Gilman, S. E., Tewksbury, J. J., & Chunco, A. J. 2011: Do species' traits predict recent shifts at expanding range edges? — *Ecology letters* 14: 677–689.

Arbabi, T., Gonzalez, J., Witt, H. H., Klein, R., & Wink, M. 2014: Mitochondrial phylogeography of the Eurasian Reed Warbler *Acrocephalus scirpaceus* and the first genetic record of *A. s. fuscus* in Central Europe. — *Ibis* 156: 799–811.

Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. 2013: RAD seq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. — *Molecular ecology* 22: 3179–3190.

Austerlitz, F., Jung-Muller, B., Godelle, B., & Gouyon, P. H. 1997: Evolution of coalescence times, genetic diversity and structure during colonization. — *Theoretical population biology* 51: 148–164.

Avilés, J. M., Stokke, B. G., Moksnes, A., Røskaft, E., Åsmul, M., & Møller, A. P. 2006: Rapid increase in cuckoo egg matching in a recently parasitized reed warbler population. — *Journal of evolutionary biology* 19: 1901–1910.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. 2008: Rapid SNP discovery and genetic mapping using sequenced RAD markers. — *PloS one* 3: e3376.

Barrett, R. D., & Schluter, D. 2008: Adaptation from standing genetic variation. — *Trends in ecology & evolution* 23: 38–44.

Bay, R. A., Rose, N., Barrett, R., Bernatchez, L., Ghalambor, C. K., Lasky, J. R., ... & Ralph, P. 2017: Predicting responses to contemporary environmental change using evolutionary response architectures. — *The American Naturalist* 189: 463–473.

Benestan, L., Moore, J. S., Sutherland, B. J., Le Luyer, J., Maaroufi, H., Rougeux, C., ... & Bernatchez, L. 2017: Sex matters in massive parallel sequencing: Evidence for biases in genetic parameter estimation and investigation of sex determination systems. — *Molecular Ecology* 26: 6767–6783.

Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., & Marron, J. S. 2004: Adjustment of systematic microarray data biases. — *Bioinformatics* 20: 105–114.

Berthouly-Salazar, C., van Rensburg, B. J., Le Roux, J. J., Van Vuuren, B. J., & Hui, C. 2012: Spatial sorting drives morphological variation in the invasive bird, *Acridotheris tristis*. — *PLoS One* 7: e38145.

Berthouly-Salazar, C., Hui, C., Blackburn, T. M., Gaboriaud, C., Van Rensburg, B. J., Van Vuuren, B. J., & Le Roux, J. J. 2013: Long-distance dispersal maximizes evolutionary potential during rapid geographic range expansion. — *Molecular ecology* 22: 5793–5804.

Bialozyt, R., Ziegenhagen, B., & Petit, R. J. 2006: Contrasting effects of long distance seed dispersal on genetic diversity during range expansion. — *Journal of evolutionary biology* 19: 12–20.

BirdLife International 2015: European Red List of Birds. http://datazone.birdlife.org/userfiles/file/Species/erlob/summarypdfs/22714722_acrocephalus_scirpaceus.pdf (accessed 10.5.2021)

BirdLife International 2021: Species factsheet: *Acrocephalus scirpaceus*. http://www.birdlife.org/species/factsheet/common-reed-warbler-acrocephalus-scirpaceus (accessed 1.5.2021)

Bors, E. K., Herrera, S., Morris Jr, J. A., & Shank, T. M. 2019: Population genomics of rapidly invading lionfish in the Caribbean reveals signals of range expansion in the absence of spatial population structure. — *Ecology and evolution* 9: 3306–3320.

Bridle, J. R., & Vines, T. H. 2007: Limits to evolution at range margins: when and why does adaptation fail? — *Trends in ecology & evolution* 22: 140–147.

Brown, J. H., Stevens, G. C., & Kaufman, D. M. 1996: The geographic range: size, shape, boundaries, and internal structure. — *Annual review of ecology and systematics* 27: 597–623.

Cabe, P. R. 1998: The effects of founding bottlenecks on genetic variation in the European starling (*Sturnus vulgaris*) in North America. — *Heredity* 80: 519–525.

Cariou, M., Duret, L., & Charlat, S. 2016: How and how much does RAD-seq bias genetic diversity estimates? — *BMC evolutionary biology* 16: 1–8.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. 2013: Stacks: an analysis tool set for population genomics. — *Molecular ecology* 22: 3124–3140.

Chen, I. C., Hill, J. K., Ohlemüller, R., Roy, D. B., & Thomas, C. D. 2011: Rapid range shifts of species associated with high levels of climate warming. — *Science* 333: 1024–1026.

Colautti, R. I., Eckert, C. G., & Barrett, S. C. 2010: Evolutionary constraints on adaptive evolution during range expansion in an invasive plant. — *Proceedings of the Royal Society B: Biological Sciences* 277: 1799–1806.

Cramp, S., & Brooks, D. J. 1992: *Handbook of the birds of Europe, the Middle East and North Africa. The birds of the western Palearctic, vol. VI. Warblers.* — Oxford University Press.

Crawford, K. M., & Whitney, K. D. 2010: Population genetic diversity influences colonization success. — *Molecular Ecology* 19: 1253–1263.

Currat, M., & Excoffier, L. 2005: The effect of the Neolithic expansion on European molecular diversity. — *Proceedings of the Royal Society B: Biological Sciences* 272: 679–688.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & McVean, G. 2011: The variant call format and VCFtools. — *Bioinformatics* 27: 2156–2158.

Davies, N. B., & Welbergen, J. A. 2009: Social transmission of a host defense against cuckoo parasitism. — *Science* 324: 1318–1320.

Davis, M. B., Shaw, R. G., & Etterson, J. R. 2005: Evolutionary responses to changing climate. — *Ecology* 86: 1704–1714.

de Pedro, M., Riba, M., González-Martínez, S. C., Seoane, P., Bautista, R., Claros, M. G., & Mayol, M. 2021: Demography, genetic diversity and expansion load in the colonizing species *Leontodon longirostris* (Asteraceae) throughout its native range. — *Molecular Ecology* 30: 1190–1205.

Duckworth, R. A. 2008: Adaptive dispersal strategies and the dynamics of a range expansion. — *The American Naturalist* 172: S4–S17.

Engler, J. O., Secondi, J., Dawson, D. A., Elle, O., & Hochkirch, A. 2016: Range expansion and retraction along a moving contact zone has no effect on the genetic diversity of two passerine birds. — *Ecography* 39: 884–893.

Excoffier, L. 2004: Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. — *Molecular ecology* 13: 853–864.

Excoffier, L., Foll, M., & Petit, R. J. 2009: Genetic consequences of range expansions. — *Annual Review of Ecology, Evolution, and Systematics* 40: 481–501.

Fix, A. G. 1997: Gene frequency clines produced by kin-structured founder effects. — *Human biology* 69: 663–673.

Fordham, D. A., Brook, B. W., Moritz, C., & Nogués-Bravo, D. 2014: Better forecasts of range dynamics using genetic data. — *Trends in ecology & evolution* 29: 436–443.

Fransson, T., & Stolt, B-O. 2005: Migration routes of North European reed warblers *Acrocephalus scirpaceus*. — *Ornis Svecica 15*: 153–160.

Garcia-Elfring, A., Barrett, R. D. H., Combs, M., Davies, T. J., Munshi-South, J., & Millien, V. 2017: Admixture on the northern front: population genomics of range expansion in the white-footed mouse *(Peromyscus leucopus*) and secondary contact with the deer mouse (*Peromyscus maniculatus*). — *Heredity* 119: 447–458.

Garroway, C. J., Bowman, J., Holloway, G. L., Malcolm, J. R., & Wilson, P. J. 2011: The genetic signature of rapid range expansion by flying squirrels in response to contemporary climate warming. — *Global Change Biology* 17: 1760–1769.

Gassert, F., Schulte, U., Husemann, M., Ulrich, W., Rödder, D., Hochkirch, A., ... & Habel, J. C. 2013: From southern refugia to the northern range margin: genetic population structure of the common wall lizard, *Podarcis muralis*. — *Journal of Biogeography* 40: 1475–1489.

Gaston, K. J. 1998: Species-range size distributions: products of speciation, extinction and transformation. — *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 353: 219–230.

Gilbert, K. J., Sharp, N. P., Angert, A. L., Conte, G. L., Draghi, J. A., Guillaume, F., ... & Whitlock, M. C. 2017: Local adaptation interacts with expansion load during range expansion: maladaptation reduces expansion load. — *The American Naturalist* 189: 368–380.

Goh, W. W. B., Wang, W., & Wong, L. 2017: Why batch effects matter in omics data, and how to avoid them. — *Trends in biotechnology* 35: 498–507.

González-Martínez, S. C., Ridout, K., & Pannell, J. R. 2017: Range expansion compromises adaptive evolution in an outcrossing plant. — *Current Biology* 27: 2544–2551.

Greenwood, P. J. 1980: Mating systems, philopatry and dispersal in birds and mammals. — *Animal behaviour* 28: 1140–1162.

Grupstra, C. G., Coma, R., Ribes, M., Leydet, K. P., Parkinson, J. E., McDonald, K., ... & Coffroth, M. A. 2017: Evidence for coral range expansion accompanied by reduced diversity of Symbiodinium genotypes. — *Coral Reefs* 36: 981–985.

Haase, M., Höltje, H., Blahy, B., Bridge, D., Henne, E., Johansson, U. S., ... & Ornés, A. S. 2019: Shallow genetic population structure in an expanding migratory bird with high breeding site fidelity, the Western Eurasian Crane *Grus grus grus*. — *Journal of Ornithology* 160: 965–972.

Hagen, S. B., Kopatz, A., Aspi, J., Kojola, I., & Eiken, H. G. 2015: Evidence of rapid change in genetic structure and diversity during range expansion in a recovering large terrestrial carnivore. — *Proceedings of the Royal Society B: Biological Sciences* 282: 20150092.

Hannah, L., Midgley, G. F., & Millar, D. 2002: Climate change-integrated conservation strategies. — *Global Ecology and Biogeography* 11: 485–495.

Harrisson, K. A., Pavlova, A., Telonis-Scott, M., & Sunnucks, P. 2014: Using genomics to characterize evolutionary potential for conservation of wild populations. — *Evolutionary Applications* 7: 1008–1025.

Hawley, D. M., Hanley, D., Dhondt, A. A., & Lovette, I. J. 2006: Molecular evidence for a founder effect in invasive house finch (*Carpodacus mexicanus*) populations experiencing an emergent disease epidemic. — *Molecular Ecology* 15: 263–275.

Heppenheimer, E., Cosio, D. S., Brzeski, K. E., Caudill, D., Van Why, K., Chamberlain, M. J., & Hinton, J. W. 2018: Demographic history influences spatial patterns of genetic diversity in recently expanded coyote (*Canis latrans*) populations. — *Heredity* 120: 183–195.

Ikonen, I. & Hagelberg E. (eds.) 2007: *Read up on reed!* — Southwest Finland Regional Environment Centre. Vammalan Kirjapaino.

IPCC, 2014: *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. — IPCC, Geneva, Switzerland, 151 pp.

Jeschke, J. M., & Strayer, D. L. 2005: Invasion success of vertebrates in Europe and North America. — *Proceedings of the National Academy of Sciences* 102: 7198–7202.

Jombart, T. 2008: *adegenet*: a R package for the multivariate analysis of genetic markers. — *Bioinformatics* 24: 1403–1405.

Jombart T. and Ahmed I. 2011: *adegenet* 1.3-1: new tools for the analysis of genome-wide SNP data. — *Bioinformatics* 27: 3070–3071.

Järvinen, O., & Ulfstrand, S. 1980: Species turnover of a continental bird fauna: Northern Europe, 1850–1970. — *Oecologia* 46: 186–195.

Kawecki, T. J. 2000: Adaptation to marginal habitats: contrasting influence of the dispersal rate on the fate of alleles with small and large effects. — *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267: 1315–1320.

Keller, S. R., Olson, M. S., Silim, S., Schroeder, W., & Tiffin, P. 2010: Genomic diversity, population structure, and migration following rapid range expansion in the Balsam Poplar*, Populus balsamifera*. — *Molecular Ecology* 19: 1212–1226.

Klopfstein, S., Currat, M., & Excoffier, L. 2006: The fate of mutations surfing on the wave of a range expansion. — *Molecular biology and evolution* 23: 482–490.

Koskimies, P. 1981: The expansion of the Great Reed Warbler *Acrocephalus arundinaceus* into Finland. — *Ornis Fennica* 58: 151–158.

Kottler, E. J., Dickman, E. E., Sexton, J. P., Emery, N. C., & Franks, S. J. 2021: Draining the swamping hypothesis: Little evidence that gene flow reduces fitness at range edges. — *Trends in Ecology & Evolution*, in press.

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. 2012: Inference of population structure using dense haplotype data. — *PLoS Genet* 8: e1002453.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., ... & Irizarry, R. A. 2010: Tackling the widespread and critical impact of batch effects in high-throughput data. — *Nature Reviews Genetics* 11: 733–739.

Legault, G., Bitters, M. E., Hastings, A., & Melbourne, B. A. 2020: Interspecific competition slows range expansion and shapes range boundaries. — *Proceedings of the National Academy of Sciences* 117: 26854–26860.

Leivo, O. 1937: *Lampikertun, Acrocephalus s. scirpaceus* (Herm.), esiintymisestä Suomessa. — *Ornis Fennica* 14: 81–90.

Li, H. 2013: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. — arXiv preprint arXiv:1303.3997.

Liebl, A. L., & Martin, L. B. 2012: Exploratory behaviour and stressor hyper-responsiveness facilitate range expansion of an introduced songbird. — *Proceedings of the Royal Society B: Biological Sciences* 279: 4375–4381.

Lombaert, E., Estoup, A., Facon, B., Joubard, B., Grégoire, J. C., Jannin, A., ... & Guillemaud, T. 2014: Rapid increase in dispersal during range expansion in the invasive ladybird *Harmonia axyridis*. — *Journal of evolutionary biology* 27: 508–517.

Lu, Z., & Yuan, K. H. 2010: Welch's t test. — In: Salkind, N. (ed.), *Encyclopedia of research design*: 1620–1623. Thousand Oaks, CA.

Løppenthin, B. 1967: *Danish breeding birds: past and present*. — Odense Universitetsforlag *(in Danish with English summary)*.

MacLean, S. A., & Beissinger, S. R. 2017: Species' traits as predictors of range shifts under contemporary climate change: A review and meta-analysis. — *Global Change Biology* 23: 4094–4105.

Malinsky, M., Trucchi, E., Lawson, D. J., & Falush, D. 2018: RADpainter and fineRADstructure: population inference from RADseq data. — *Molecular biology and evolution* 35: 1284–1290.

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. 2007: Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. — *Genome research* 17: 240–248.

Montadert, M., & Leonard, P. 2006: Post-juvenile dispersal of Hazel Grouse *Bonasa bonasia* in an expanding population of the southeastern French Alps. — *Ibis* 148: 1–13.

Nadeau, C. P., & Urban, M. C. 2019: Eco-evolution on the edge during climate change. — *Ecography* 42: 1280–1297.

Nazareno, A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. G. 2017: Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. — *Molecular Ecology Resources* 17: 1136–1147.

Neel, J. V. 1973: "Private" genetic variants and the frequency of mutation among South American Indians. — *Proceedings of the National Academy of Sciences* 70: 3311–3315.

Nei, M. 1973: Analysis of gene diversity in subdivided populations. — *Proceedings of the National Academy of Sciences* 70: 3321–3323.

Nei, M., & Li, W. H. 1979: Mathematical model for studying genetic variation in terms of restriction endonucleases. — *Proceedings of the National Academy of Sciences* 76: 5269–5273.

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. 2018: These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. — *Molecular Ecology* 27: 3193–3206.

Paradis, E., Baillie, S. R., Sutherland, W. J., & Gregory, R. D. 1998: Patterns of natal and breeding dispersal in birds. — *Journal of Animal ecology* 67: 518–536.

Parmesan, C., & Yohe, G. 2003: A globally coherent fingerprint of climate change impacts across natural systems. — *Nature* 421: 37–42.

Pecl, G. T., Araújo, M. B., Bell, J. D., Blanchard, J., Bonebrake, T. C., Chen, I. C., ... & Williams, S. E. 2017: Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. — *Science* 355 (art. eaai9214).

Peischl, S., Kirkpatrick, M., & Excoffier, L. 2015: Expansion load and the evolutionary dynamics of a species range. — *The American Naturalist* 185: E81–E93.

Pereira, H. M., Leadley, P. W., Proença, V., Alkemade, R., Scharlemann, J. P., Fernandez-Manjarrés, J. F., ... & Walpole, M. 2010: Scenarios for global biodiversity in the 21st century. — *Science* 330: 1496–1501.

Pfennig, K. S., Kelly, A. L., & Pierce, A. A. 2016: Hybridization as a facilitator of species range expansion. — *Proceedings of the Royal Society B: Biological Sciences* 283: 20161329.

Pfenninger, M., Nowak, C., & Magnin, F. 2007: Intraspecific range dynamics and niche evolution in *Candidula* land snail species. — *Biological Journal of the Linnean Society* 90: 303–317.

Pierce, A. A., Zalucki, M. P., Bangura, M., Udawatta, M., Kronforst, M. R., Altizer, S., ... & de Roode, J. C. 2014: Serial founder effects and genetic differentiation during worldwide range expansion of monarch butterflies. — *Proceedings of the Royal Society B: Biological Sciences* 281: 20142230.

Pigot, A. L., Owens, I. P., & Orme, C. D. L. 2010: The environmental limits to geographic range expansion in birds. — *Ecology Letters* 13: 705–715.

Polechová, J. 2018: Is the sky the limit? On the expansion threshold of a species' range. — *PLoS biology* 16: e2005372.

Procházka, P., Stokke, B. G., Jensen, H., Fainová, D., Bellinvia, E., Fossøy, F., ... & Soler, M. 2011: Low genetic differentiation among reed warbler *Acrocephalus scirpaceus* populations across Europe. — *Journal of Avian Biology* 42: 103–113.

Pruett, C., & Winker, K. 2008: The effects of sample size on population genetic diversity estimates in song sparrows *Melospiza melodia*. — *Journal of Avian Biology* 39: 252–256.

Pujol, B., & Pannell, J. R. 2008: Reduced responses to selection after species range expansion. — *Science* 321: 96.

Ramos, R., Song, G., Navarro, J., Zhang, R., Symes, C. T., Forero, M. G., & Lei, F. 2016: Population genetic structure and long-distance dispersal of a recently expanding migratory bird. — *Molecular phylogenetics and evolution* 99: 194–203.

Ray, N., Currat, M., & Excoffier, L. 2003: Intra-deme molecular diversity in spatially expanding populations. — *Molecular biology and evolution* 20: 76–86.

Refsnider, J. M., & Janzen, F. J. 2012: Behavioural plasticity may compensate for climate change in a long-lived reptile with temperature-dependent sex determination. — *Biological Conservation* 152: 90–95.

Regier, A. A., Farjoun, Y., Larson, D. E., Krasheninina, O., Kang, H. M., Howrigan, D. P., ... & Hall, I. M. 2018: Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. — *Nature communications* 9: 1–8.

Rendine, S., Piazza, A., & Cavalli-Sforza, L. L. 1986: Simulation and separation by principal components of multiple demic expansions in Europe. — *The American Naturalist* 128: 681–706.

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. 2019: Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. — *Molecular Ecology* 28: 4737–4754.

Romeiks, C. 2012: European reed warbler. [Photograph] https://commons.wikimedia.org/wiki/File:Acrocephalus_scirpaceus_vogelartinfo_chris_romeiks_R7F6188.jpg (accessed 11.5.2021)

Roques, L., Garnier, J., Hamel, F., & Klein, E. K. 2012: Allee effect promotes diversity in traveling waves of colonization. — *Proceedings of the National Academy of Sciences* 109: 8828–8833.

Rózsa, J., Strand, T. M., Montadert, M., Kozma, R., & Höglund, J. 2016: Effects of a range expansion on adaptive and neutral genetic diversity in dispersal limited Hazel grouse (*Bonasa bonasia*) in the French Alps. — *Conservation genetics* 17: 401–412.

Røed, T. 1994: Rørsanger — In: *Norsk Fugleatlas*: 382–83. Norsk Ornitologisk Forening. (Available online: https://www.birdlife.no/fuglekunnskap/fugleatlas/pdf/rorsanger.pdf)

Savolainen, O., Lascoux, M., & Merilä, J. 2013: Ecological genomics of local adaptation. — *Nature Reviews Genetics* 14: 807–820.

Schrey, A. W., Grispo, M., Awad, M., Cook, M. B., McCoy, E. D., Mushinsky, H. R., ... & Martin, L. B. 2011: Broad-scale latitudinal patterns of genetic diversity among native European and introduced house sparrow (*Passer domesticus*) populations. — *Molecular Ecology* 20: 1133–1143.

Sebastiani, P., Solovieff, N., Puca, A., Hartley, S. W., Melista, E., Dworkis, D. A., … Perls, T. T. 2011: Retraction. — *Science* 333: 404.

Sexton, J. P., McIntyre, P. J., Angert, A. L., & Rice, K. J. 2009: Evolution and ecology of species range limits. — *Annual Review of Ecology, Evolution and Systematics* 40: 415–436.

Sexton, J. P., Hangartner, S. B., & Hoffmann, A. A. 2014: Genetic isolation by environment or distance: which pattern of gene flow is most common? — *Evolution* 68: 1–15.

Short, K. H., & Petren, K. 2011: Fine-scale genetic structure arises during range expansion of an invasive gecko. — *PLoS one* 6: e26258.

Simberloff, D., Souza, L., Nuñez, M. A., Barrios-Garcia, M. N., & Bunn, W. 2012: The natives are restless, but not often and mostly when disturbed. — *Ecology* 93: 598–607.

Skellam, J. G. 1951: Random dispersal in theoretical populations. — *Biometrika* 38: 196–218.

Slatkin, M. 1985: Gene flow in natural populations. — *Annual review of ecology and systematics* 16: 393–430.

Song, G., Yu, L., Gao, B., Zhang, R., Qu, Y., Lambert, D. M., ... & Lei, F. 2013: Gene flow maintains genetic diversity and colonization potential in recently range-expanded populations of an Oriental bird, the Light-vented Bulbul (*Pycnonotus sinensis*, Aves: Pycnonotidae). — *Diversity and Distributions* 19: 1248–1262.

Sutter, M., & Kawecki, T. J. 2009: Influence of learning on range expansion and adaptation to novel habitats. — *Journal of Evolutionary Biology* 22: 2201–2214.

Swaegers, J., Mergeay, J., Therry, L., Larmuseau, M. H. D., Bonte, D., & Stoks, R. 2013: Rapid range expansion increases genetic differentiation while causing limited reduction in genetic diversity in a damselfly. — *Heredity* 111: 422–429.

Szűcs, M., Melbourne, B. A., Tuff, T., & Hufbauer, R. A. 2014: The roles of demography and genetics in the early stages of colonization. — *Proceedings of the Royal Society B: Biological Sciences* 281: 20141073.

Szűcs, M., Vahsen, M. L., Melbourne, B. A., Hoover, C., Weiss-Lehman, C., & Hufbauer, R. A. 2017: Rapid adaptive evolution in novel environments acts as an architect of population range expansion. — *Proceedings of the National Academy of Sciences* 114: 13501–13506.

Thorogood, R., & Davies, N. B. 2013: Reed warbler hosts fine-tune their defenses to track three decades of cuckoo decline. — *Evolution* 67: 3545–3555.

Travis, J. M., & Dytham, C. 2002: Dispersal evolution during invasions. — *Evolutionary Ecology Research* 4: 1119–1129.

Tuomainen, U., & Candolin, U. 2011: Behavioural responses to human-induced environmental change. — *Biological Reviews* 86: 640–657.

Tylianakis, J. M., Laliberté, E., Nielsen, A., & Bascompte, J. 2010. Conservation of species interaction networks. — *Biological conservation* 143: 2270–2279.

Valkama, Jari, Vepsäläinen, Ville & Lehikoinen, Aleksi 2011*: The Third Finnish Breeding Bird Atlas.* — Finnish Museum of Natural History and Ministry of Environment. http://atlas3.lintuatlas.fi/english (cited 27.9.2020)

Wagner, N. K., Ochocki, B. M., Crawford, K. M., Compagnoni, A., & Miller, T. E. 2017: Genetic mixture of multiple source populations accelerates invasive range expansion. — *Journal of Animal Ecology* 86: 21–34.

Wallingford, P. D., Morelli, T. L., Allen, J. M., Beaury, E. M., Blumenthal, D. M., Bradley, B. A., ... & Sorte, C. J. 2020: Adjusting the lens of invasion biology to focus on the impacts of climate-driven range shifts. — *Nature Climate Change* 10: 398–405.

Waters, C. N., Zalasiewicz, J., Summerhayes, C., Barnosky, A. D., Poirier, C., Gałuszka, A., ... & Wolfe, A. P. 2016: The Anthropocene is functionally and stratigraphically distinct from the Holocene. — *Science* 351 (art. aad2622).

Wegmann, D., Currat, M., & Excoffier, L. 2006: Molecular diversity after a range expansion in heterogeneous environments. — *Genetics* 174: 2009–2020.

Welles, S.R. & Dlugosch, K. M. 2019: Population genomics of colonization and invasion. — In: Rajora, O.P. (ed.), *Population genomics: concepts, approaches and applications*: 655–683. Springer.

Wennersten, L., & Forsman, A. 2012: Population-level consequences of polymorphism, plasticity and randomized phenotype switching: A review of predictions. — *Biological Reviews* 87: 756–767.

White, T. A., Perkins, S. E., Heckel, G., & Searle, J. B. 2013: Adaptive evolution during an ongoing range expansion: the invasive bank vole (*Myodes glareolus*) in Ireland. — *Molecular Ecology* 22: 2971–2985.

Whitlock, M. C., & McCauley, D. E. 1990: Some population genetic consequences of colony formation and extinction: genetic correlations within founding groups. — *Evolution* 44: 1717–1724.

Wickham, H. 2011: ggplot2. — *Wiley Interdisciplinary Reviews: Computational Statistics* 3: 180–185.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. 2019: Welcome to the Tidyverse. — *Journal of Open Source Software* 4: 1686.

Wikström, D. 1945: Lampikerttu, *Acrocephalus s. scirpaceus* (Herm.) pesinyt Suomessa jo v. 1922. — *Ornis Fennica* 22: 29–31.

Wright, S. 1931: Evolution in Mendelian Populations. — *Genetics* 16: 97–159.

Wright, S. 1943: Isolation by distance. — *Genetics* 28: 114.

Appendix A. Genetic diversity analysis: assessing the homogeneity of the range core (Czech and Slovakian) samples

Settings: Stacks *populations*:

-p 2, -r 0.80, -H, --min-mac 3, --max-obs-het 0.70, --blacklist (excluding the sex chromosome and all loci that contain SNPs with mean depth lower than 15 or higher than 160, which is approximately twice the mean depth per site)

fineRADstructure:

-n 5

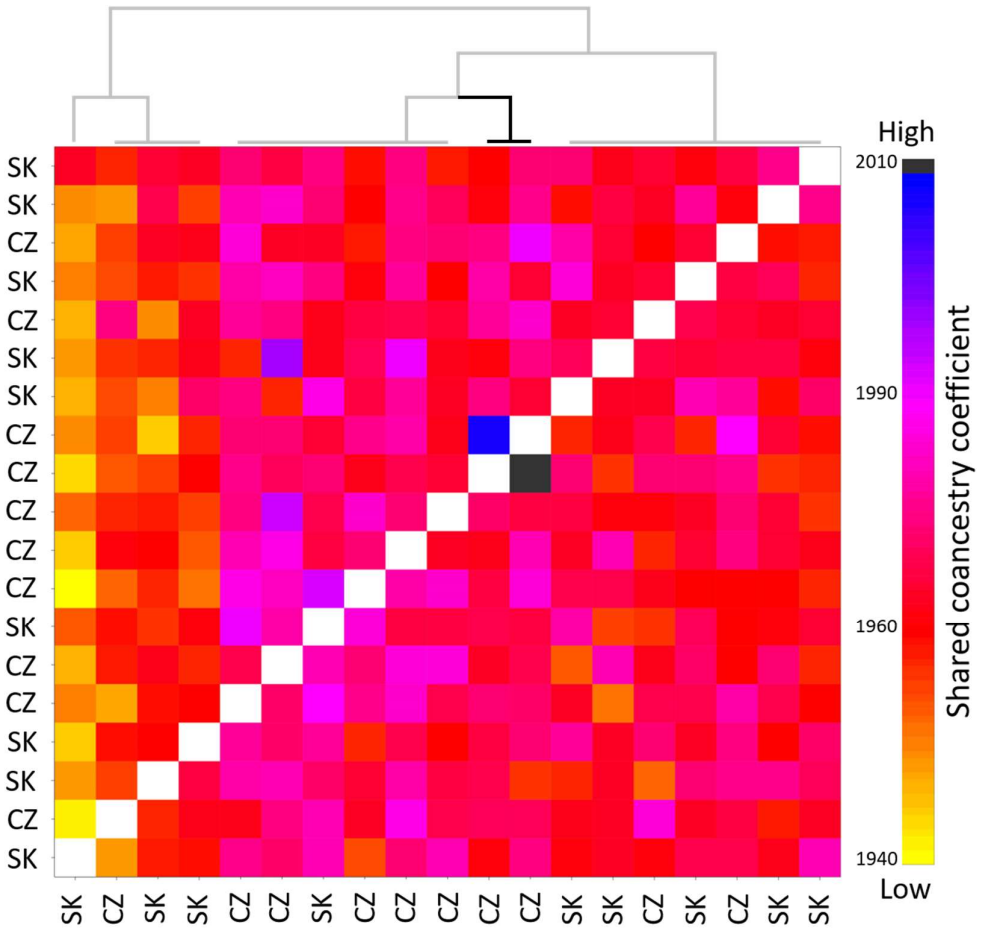


Figure 8. No distinct genetic clustering in the fineRADstructure plot. Black branches indicate higher than 95 % posterior probability support. CZ = Czech Republic, SK = Slovakia. Based on this result, samples were treated as a single population in the genetic diversity analysis.

Appendix B. Comparison of population genetic statistics between sequencing batches

Table 4. Population genetic statistics for the replicate samples before normalizing the mean coverage, calculated using the Stacks *populations* program. N = number of sequenced samples, H_o = observed heterozygosity, H_e = expected heterozygosity, π = nucleotide diversity, F_{IS} = inbreeding coefficient. All formulas used for calculations in Catchen et al. 2013.

	N	Variant sites	Mean coverage of variant sites	Polymorphic sites	Private alleles	H_o	Var(H_o)	H_e	Var(H_e)	π	Var(π)	F_{IS}	Var(F_{IS})
BATCH 1	5	21951	100.59	21928	134	0.269	0.038	0.348	0.012	0.386	0.015	0.272	0.203
BATCH 2	5	21951	53.05	21817	23	0.142	0.027	0.362	0.010	0.402	0.012	0.620	0.189

Table 5. Population genetic statistics for the replicate samples after normalizing the mean coverage. The normalization did not have a notable effect on the values of the population genetic statistics, and did not therefore reduce the batch effect. N = number of sequenced samples, H_o = observed heterozygosity, H_e = expected heterozygosity, π = nucleotide diversity, F_{IS} = inbreeding coefficient. All formulas used for calculations in Catchen et al. 2013.

	N	Variant sites	Mean coverage of variant sites	Polymorphic sites	Private alleles	H_o	Var(H_o)	H_e	Var(H_e)	π	Var(π)	F_{IS}	Var(F_{IS})
BATCH 1	5	19787	39.84	19767	132	0.260	0.038	0.348	0.012	0.386	0.015	0.293	0.209
BATCH 2	5	19787	38.81	19655	20	0.138	0.027	0.362	0.010	0.402	0.012	0.629	0.188

Appendix C. Results from fineRADstructure analysis in the range edge population (Finland) after increasing the SNP threshold to max 10 SNPs per locus

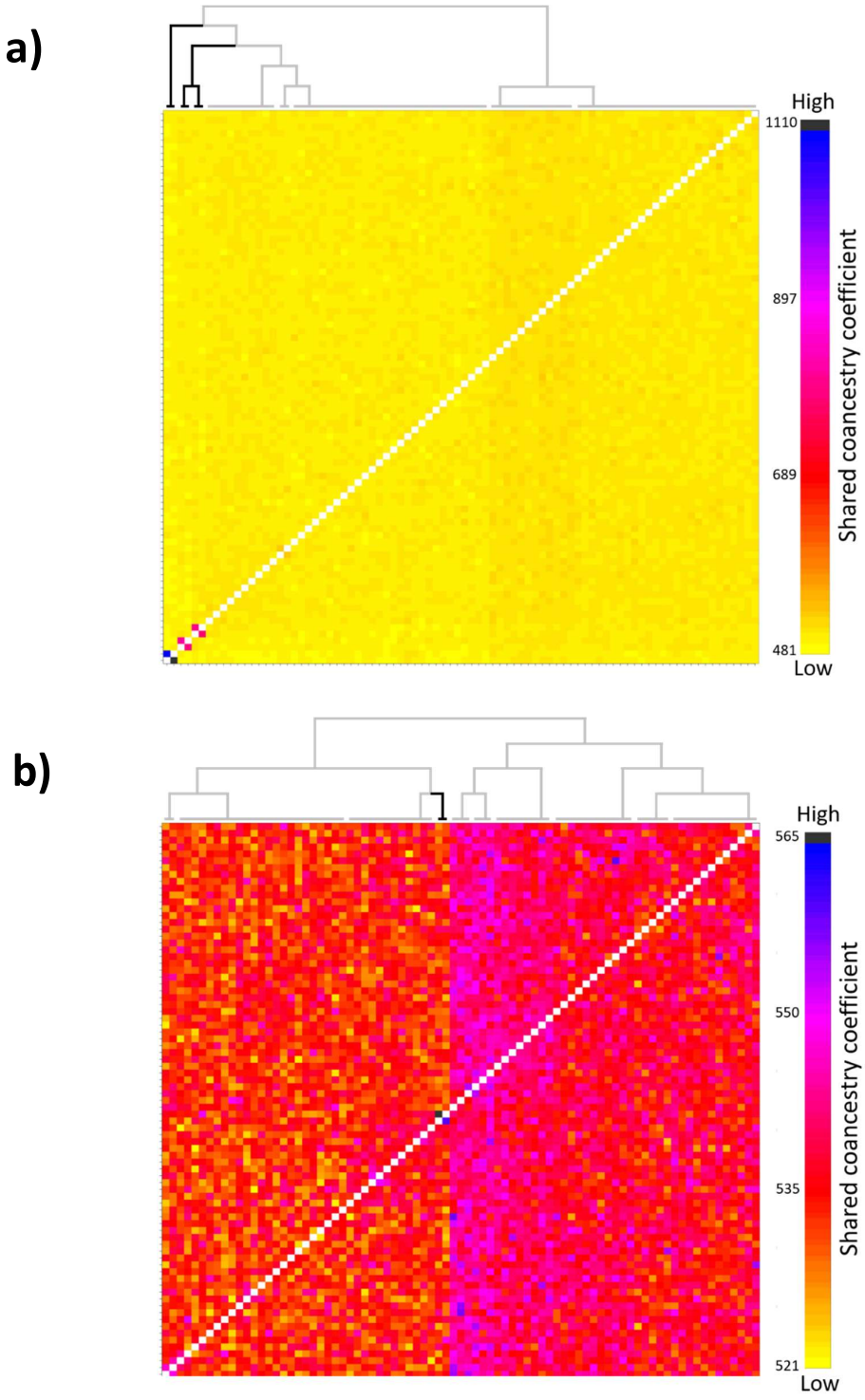
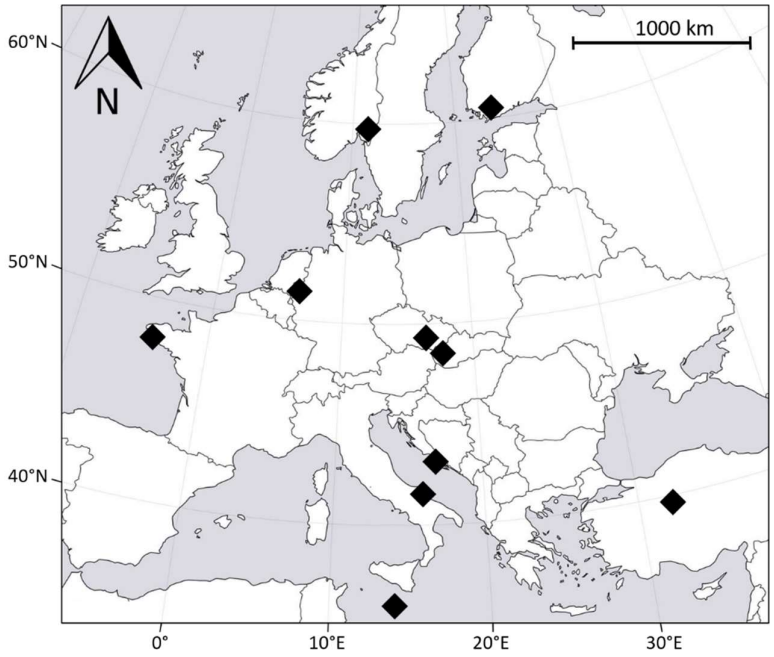


Figure 9. FineRADstructure results of **a)** all range edge samples (41 600 loci), **b)** samples after removal of one individual from each outlier pair (42 900 loci). Black branches indicate > 95 % posterior probability.

Appendix D. Population structure across the European range of the reed warbler

The population structure analyses were performed also for all European populations that were included in the two sequencing batches (Fig. 10). This was done to test the sufficient resolution of the analyses in detecting low levels of population structure, and to see how the recently established populations (Finland and Norway) cluster at a wider spatial scale. A population map with a maximum of 10 individuals from each ten sampling countries was created (n = 94), treating each country as a separate population. Countries with more than 10 sequenced individuals were downsampled otherwise randomly, but checking that the chosen individuals did not have exceptionally high amounts of missing data.



Location	N	Batch
Croatia	9	B1
Czech R.	10	B1
Finland	10	B2
France	8	B1
Germany	10	B1
Italy	10	B1
Malta	10	B1
Norway	10	B1
Slovakia	10	B1
Turkey	7	B1

Figure 10. Sampling locations across the European range. The number of sampling sites within each location, the number of sequenced samples, and the batch in which the samples were sequenced are shown in the table.

For the PCA, the analysis was restricted to one random SNP per locus. *Populations* was run with the following settings: -p 10, -r 0.80, --min-mac 3, --max-obs-het 0.70, --write-random-snp, --blacklist (sex chromosome, loci that contain sites with lower mean depth than 15 and higher mean depth than 150). The PCA was performed similarly as with the Finnish population in section 2.5.1. The same population map was used to create the data set for the fineRADstructure analysis. *Populations* was run with the same settings as

for the PCA, with the exception of retaining all SNPs within each locus, and applying the -p and -r filters haplotype-wise (-H). The fineRADstructure program was run as with the Finnish population in section 2.5.1, except for performing the analysis only once with a threshold of 5 SNPs allowed per locus (-n 5).

Results from the population structure analyses

The results including outlier individuals can be seen in Fig. 11. After removing these outliers (one Slovakian and all Maltan samples) to get a closer view of the population structure, the data set for PCA consisted of 41 181 unlinked SNPs (n = 83, missingness per individual 0.5 – 12.2 %). The resulting plot (Fig. 12 a) shows spatial clustering on the PC1 axis, with the most northern populations at one end of the gradient, and the most southern ones roughly at the other end. This principal component stands out in the eigenvalue plot, explaining distinctly most variation in the data (%). Interestingly, the PC2 axis separates the Finnish and Norwegian samples into a single cluster, which is not directly on the spatial continuum with the other populations. PC3 picks up two pairs of individuals with higher pairwise genetic similarity than the other samples: one pair from Germany and one from Italy.

The fineRADstructure analysis was run with a haplotype dataset from ca 27 600 loci (missingness per individual 0.6 – 12.1 %). The resulting tree and coancestry matrix (Fig. 12 b) divide the samples into two main clusters: the Finnish and Norwegian samples in one cluster (on the right in the cladogram), and Central and Southern European samples in the other one. The Central and Southern European cluster is further divided into two well-supported groups (posterior probability > 95 %, denoted by black branches in the cladogram), and these groups seem to match the known migratory divide of European reed warblers (Procházka et al. 2011). Like PCA, also fineRADstructure detected a few pairs of individuals with higher than average pairwise coancestry (black, blue or purple in the heat map). The clustering of the northern range edge populations, Finland and Norway, suggests genetic differentiation during the northward range expansion of the species despite the high levels of retained genetic diversity (section 3.3). Further studies are needed for determining whether this differentiation is more pronounced than expected just by isolation-by-distance (IBD) (Wright 1943).

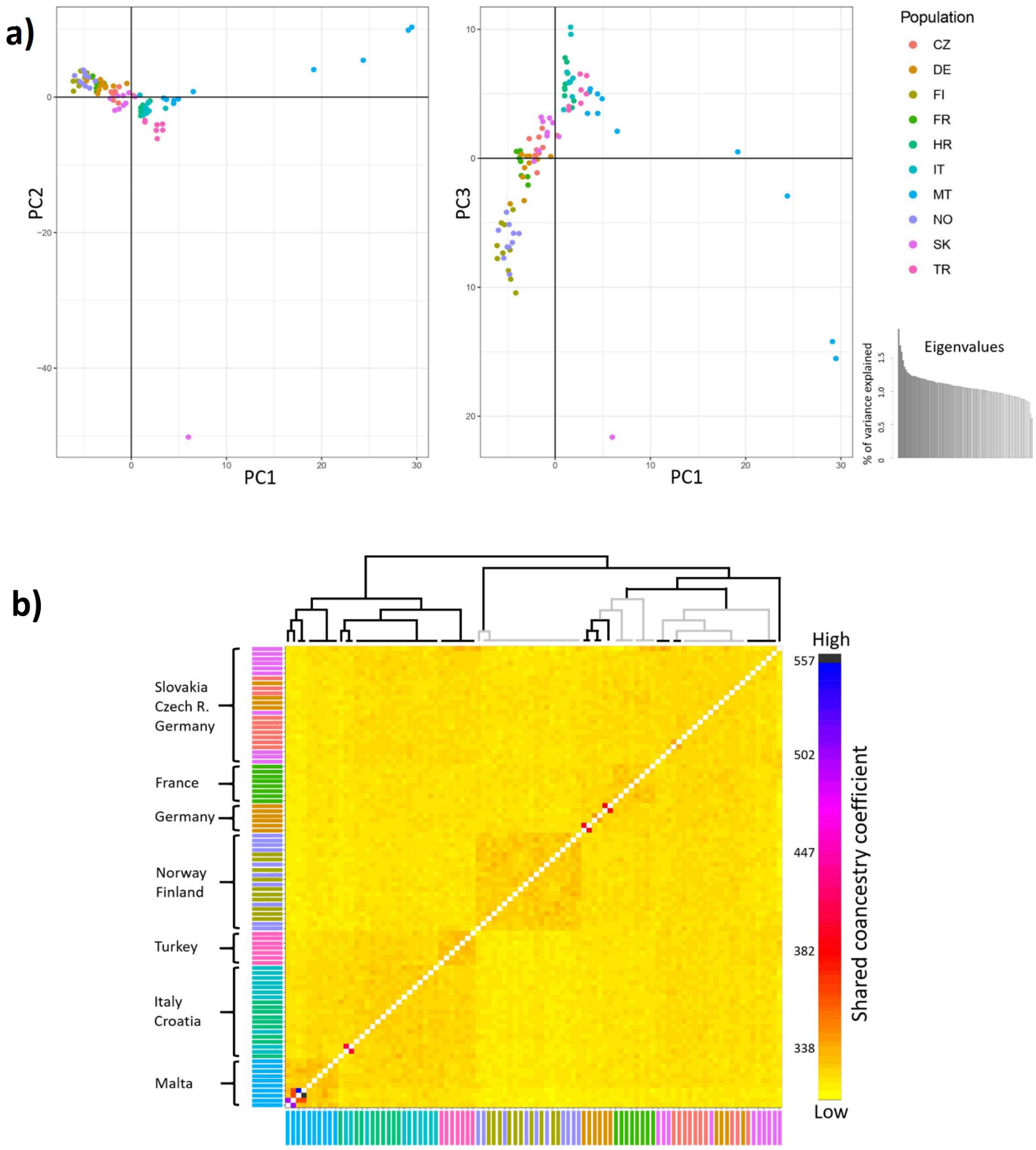


Figure 11. a) PCA of the sampled European populations (42 818 SNPs) before removing outliers. The scatterplots show factor scores of individuals in first and second (left) and first and third principal components (right). The percentage of explained variation by each PC is shown in the eigenvalue plot. **b)** FineRADstructure analysis results (29 000 loci) before removing outliers. Black branches indicate > 95 % posterior probability support.

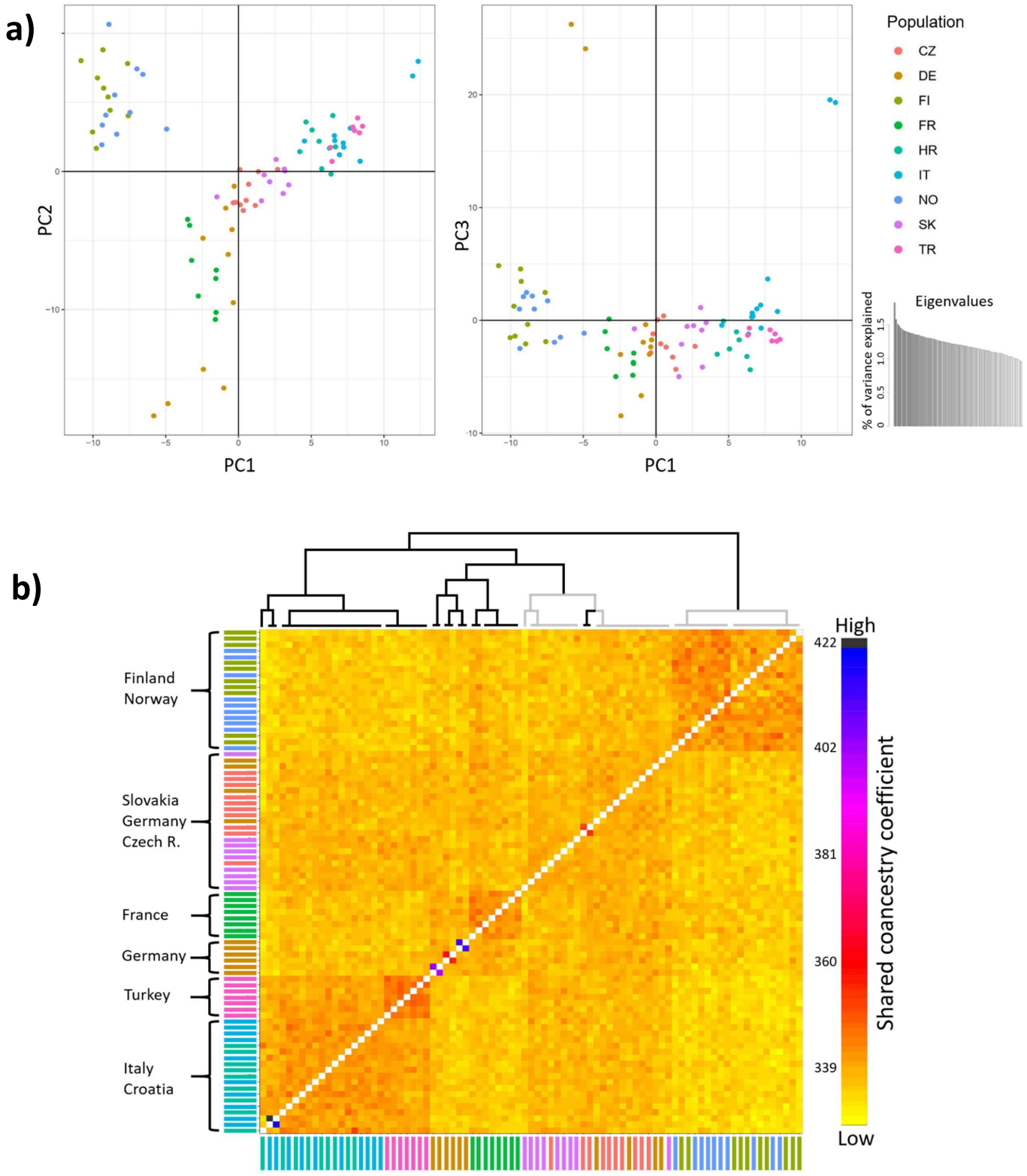


Figure 12. Spatial grouping of the reed warbler individuals across the sampled European range, after the removal of outliers. **a)** Principal component analysis (PCA) of the genomic data (41 181 SNPs). The scatterplots show factor scores of individuals in first and second (left) and first and third principal components (right). The percentage of explained variation by each PC is shown in the eigenvalue plot. The

individuals are coloured according to their sampling population (country). **b)** Output of fineRADstructure analysis (27 600 RAD loci) after removing outliers. In the population tree, branches with higher than 95 % posterior probability are shown in black, grey branches indicate support below this limit. The heat map indicates pairwise coancestry between individuals: lowest coancestry estimates in the data are represented by yellow, intermediate by red, and highest by blue and black. Sample individuals are illustrated as small bars on the left side and under the heat map: the colour of each bar represents the sampling country of the individual.